FUNDAMENTALS OF MATHEMATICAL STATISTICS

(A Modern Approach)

A Textbook written completely on modern lines for Degree, Honours, Post-graduate Students of all Indian Universities and Indian Civil Services, Indian Statistical Service Examinations.

(Contains, besides complete theory, more than 650 fully solved examples and more than 1,500 thought-provoking Problems with Answers, and Objective Type Questions)

S.C. GUPTA
Reader in Statistics
Hindu College,
University of Delhi
Delhi

V.K. KAPOOR
Reader in Mathematics
Shri Ram College of Commerce
University of Delhi
Delhi

Tenth Revised Edition
(Greatly Improved)

SULTAN CHAND & SONS
Educational Publishers
New Delhi

* First Edition : Sept. 1970
Tenth Revised Edition : August 2000
Reprint : 2002

* Price : **Rs. 210.00**

ISBN 81-7014-791-3

* *Published by* :
Sultan Chand & Sons
23, Darya Ganj, New Delhi-110002
Phones : 3277843, 3266105, 3281876

* *Laser typeset by* : **T.P.**

Printed at:- New A.S. Offset Press Laxmi Nagar Delhi-92

Dedicated to
Our Teacher
Professor H.C. Gupta
Who Initiated
The Teaching of
Mathematical Statistics
At the
University of Delhi

PREFACE

TO THE TENTH EDITION

The book has been revised keeping in mind the comments and suggestions received from the readers. An attempt is made to eliminate the misprints/errors in the last edition. Further suggestions and criticism for the improvement of the book will be most welcome and thankfully acknowledged.

August 2000 **S.C. GUPTA**
V.K. KAPOOR

TO THE NINTH EDITION

The book originally written twenty-four years ago has, during the intervening period, been revised and reprinted several times. The authors have, however, been thinking, for the last few years that the book needed not only a thorough revision but rather a complete rewriting. They now take great pleasure in presenting to the readers the ninth completely revised and enlarged edition of the book. The subject-matter in the whole book has been rewritten in the light of numerous criticisms and suggestions received from the users of the previous editions in India and abroad.

Some salient features of the new edition are:

- *The entire text, especially Chapter 5 (Random Variables), Chapter 6 (Mathematical Expectation), Chapters 7 and 8 (Theoretical Discrete and Continuous Distributions), Chapter 10 (Correlation and Regression), Chapter 15 (Theory of Estimation), has been restructured, rewritten and updated to cater to the revised syllabi of Indian universities, Indian Civil Services and various other competitive examinations.*
- *During the course of rewriting, it has been specially borne in mind to retain all the basic features of the previous editions especially the simplicity of presentation, lucidity of style and analytical approach which have been appreciated by teachers and students all over India and abroad.*
- *A number of typical problems have been added as solved examples in each chapter. These will enable the reader to have a better and thoughtful understanding of the basic concepts of the theory and its various applications.*
- *Several new topics have been added at appropriate places to make the treatment more comprehensive and complete. Some of the obvious ADDITIONS are:*

§ 8·1.5 Triangular Distribution p. 8·10 to 8·12
§ 8·8.3 Logistic Distribution p. 8·92 to 8·95
§ 8·10 Remarks 2, Convergence in Distribution of Law p. 8·106
§ 8·10.3. Remark 3, Relation between Central Limit Theorem and Weak Law of Large Numbers p. 8·110
§ 8·10.4 Cramer's Theorem p . 8·111-8.112, 8·114-8·115 — Example 8.46

§ 8·14 to § 8·14·6	Order Statistics — Theory, Illustrations and Exercise Set p. 8·136 to 8·151
§ 8·15	Truncated Distributions—with Illustrations p. 8·151 to 8·156
§ 10·6·1	Derivation of Rank Correlation Formula for Tied Ranks p. 10·40-10·41
§ 10·7·1	Lines of Regression—Derivation (Aliter) p. 10·50-10·51. Example 10·21 p. 10·55
§ 10·10·2	Remark to § 10·10·2 — Marginal Distributions of Bivariate Normal Distribution p. 10·88-10·90 Theorem 10·5, p. 10·86. and Theorem 10·6, p. 10·87 on Bivariate Normal Distribution. Solved Examples 10·31, 10·32, pages 10·96-10·97 on BVN Distribution. Theorem 13·5 Alternative Proof of Distribution of $(\bar{X}, s^2)$ using m.g.f. p.13·19 to 13·21
§ 13·11	χ^2-Test for pooling of Probabilities (P_λ Test) p. 13·69
§ 15·4·1	Invariance property of Consistent Estimators—Theorem 15·1, pp 15·3
§ 15·4·2	Sufficient Conditions for Consistency—Theorem 15·2, p. 15·3
§ 15·5·5	MVUE : Theorem 15·4, p. 15·12-15·13
§ 15·7	Remark 1. Minimum Variance Bound (MVB), Estimator, p.15·24
§ 15·7·1	Conditions for the equality sign in Cramer-Rao (CR) inequality, p. 15·25 to 15·27
§ 15·8	Complete family of Distributions (with illustrations), p. 15·31 to 15·34 Theorem 15·10 (Blackwellisation), p. 15·36. Theorems 15·16 and 15·17 on MLE, p. 15·55.
§ 16·5·1	Unbiased Test and Unbiased Critical Region. Theorem 16·2-pages 16·9-16·10
§ 16·5·2	Optimum Regions and Sufficient Statistics, p.16·10-16·11 Remark to Example 16·6, p. 16·17-16·18 and Remarks 1, 2 to Example 16·7, p. 16·20 to 16·22; Graphical Representation of Critical Regions.

• Exercise sets at the end of each chapter are substantially reorganised. Many new problems are included in the exercise sets. Repetition of questions of the same type (more than what is necessary) has been avoided. Further in the set of exercises, the problems have been carefully arranged and properly graded. More difficult problems are put in the miscellaneous exercise at the end of each chapter.

• Solved examples and unsolved problems in the exercise sets have been drawn from the latest examination papers of various Indian Universities, Indian Civil Services, etc.

• *An attempt has been made to rectify the errors in the previous editions.*

• ***The present edition incorporates modern viewpoints. In fact with the addition of new topics, rewriting and revision of many others and restructuring of exercise sets, altogether a new book, covering the revised syllabi of almost all the Indian universities, is being presented to the reader. It is earnestly hoped that, in the new form, the book will prove of much greater utility to the students as well as teachers of the subject.***

We express our deep sense of gratitude to our Publishers M/s Sultan Chand & Sons and printers DRO Phototypesetter for their untiring efforts, unfailing courtesy, and co-operation in bringing out the book, in such an elegant form. We are also thankful to our several colleagues, friends and students for their suggestions and encouragement during the preparing of this revised edition.

Suggestions and criticism for further improvement of the book as well as intimation of errors and misprints will be most gratefully received and duly acknowledged.

S C. GUPTA & V.K. KAPOOR

TO THE FIRST EDITION

Although there are a large number of books available covering various aspects in the field of Mathematical Statistics, there is no comprehensive book dealing with the various topics on Mathematical Statistics for the students. The present book is a modest though determined bid to meet the requirements of the students of Mathematical Statistics at Degree, Honours and Post-graduate levels. The book will also be found of use by the students preparing for various competitive examinations. While writing this book our goal has been to present a clear, interesting, systematic and thoroughly teachable treatment of Mathematical Statistics and to provide a textbook which should not only serve as an introduction to the study of Mathematical Statistics but also carry the student on to such a level that he can read with profit the numercus special monographs which are available on the subject. In any branch of Mathematics, it is certainly the teacher who holds the key to successful learning, Our aim in writing this book has been simply to assist the teacher in conveying to the students more effectively a thorough understanding of Mathematical Statistics.

The book contains sixteen chapters (equally divided between two volumes). The first chapter is devoted to a concise and logical development of the subject. The second and third chapters deal with the frequency distributions, and measures of average and dispersion. Mathematical treatment has been given to the proofs of various articles included in these chapters in a very logical and simple manner. The theory of probability which has been developed by the application of the set theory has been discussed quite in detail. A large number of theorems have been deduced using the simple tools of set theory. The *simple applications of probability are also given. The chapters on mathematical expectation and theoretical distributions (discrete as well as continuous) have been written keeping the latest ideas in mind. A new treatment has been given to the chapters on correlation, regression and bivariate normal distribution using the concepts of mathematical expectation. The thirteenth and fourteenth chapters deal mainly with the various sampling distributions and the various tests of significance which can be derived from them. In chapter 15, we have discussed concisely statistical inference (estimation and testing of hypothesis). Abundant material is given in the chapter on finite differences and numerical integration. The whole of the relevant theory is arranged in the form of serialised articles which are concise and to the point without being insufficient. The more difficult sections will, in general, be found towards the end of each chapter. We have tried our best to present the subject so as to be within the easy grasp of students with varying degrees of intellectual attainment.*

Due care has been taken of the examination needs of the students and, wherever possible, indication of the year, when the articles and problems were set in the examination as been given. While writing this text, we have gone through the syllabi and examination papers of almost all Indian universities where the subject is taught so as to make it as comprehensive as possible. Each chapter contains a large number of carefully graded worked problems mostly drawn from university papers with a view to acquainting the student with the typical questions pertaining to each topic. Furthermore, to assist the student to gain proficiency in the subject, a large number of properly graded problems mainly drawn from examination papers of various universities are given at the end of each chapter. The questions and problems given at the end of each chapter usually require for their solution a thoughtful use of concepts. During the preparation of the text we have gone through a vast body of literature available on the subject, a list of which is given at the end of the book. It is expected that the bibliography given at the end of the book will considerably help those who want to make a detailed study of the subject

The lucidity of style and simplicity of expression have been our twin objects to remove the awe which is usually associated with most mathematical and statistical textbooks.

While every effort has been made to avoid printing and other mistakes, we crave for the indulgence of the readers for the errors that might have inadvertently crept in. We shall consider our efforts amply rewarded if those for whom the book is intended are benefited by it. Suggestions for the improvement of the book will be highly appreciated and will be duly incorporated.

SEPTEMBER 10, 1970 **S.C. GUPTA & V.K. KAPOOR**

Contents

Chapter - 7

Theoretical Discrete Distributions **7·1 — 7·114**

Chapter — 11
Theory of Attributes **11·1 — 11·22**

Chapter — 12
Sampling and Large Sample Tests **12·1 — 12·50**

Chapter — 14

Exact Sampling Distributions (Continued) (t, F and z distributions)

Fundamentals of Mathematical Statistics

S.C. GUPTA V.K. KAPOOR
Hindu College, University of Delhi, Delhi — *Shri Ram College of Commerce University of Delhi, Delhi*

Tenth Edition 2002	Pages *xx* + 1284	22 × 14 cm	ISBN 81-7014-791-3	**Rs 210.00**

Special Features

- Comprehensive and analytical treatment is given of all the topics.
- Difficult mathematical deductions have been treated logically and in a very simple manner.
- It conforms to the latest syllabi of the Degree and Post graduate examinations in Mathematics, Statistics and Economics.

Contents

Introduction • Frequency Distribution and Measures of Central Tendency • Measures of Dispersion, Skewness and Kurtosis • Theory of Probability • Random Variables—Distribution Function • Mathematical Expectation, Generating Functions and Law of Large Numbers • Theoretical Discrete Distributions • Theoretical Continuous Distributions • Curve Fitting and Principle of Least Squares • Correlation, Regression, Bivariate Normal Distribution and Partial & Multiple Correlation • Theory of Attributes • Sampling and Large Sample Tests of Mean and Proportion • Sampling Distribution Exact (Chi-square Distribution) • Exact Sampling Distributions (t, F and Z Distributions) • Theory of Estimation • Testing of Hypothesis, Sequential Analysis and Non-parametric Methods.

Elements of Mathematical Statistics

S.C. GUPTA V.K. KAPOOR

Third Edition 2001	Pages *xiv* + 489	ISBN 81-7014-291-1	**Rs 70.00**

- Prepared specially for B.Sc. students, studying Statistics as subsidiary or ancillary subject.

Contents

Introduction—Meaning and Scope • Frequency Distributions and Measures of Central Tendency • Measures of Dispersion, Skewness and Kurtosis • Theory of Probability • Random Variables—Distribution Functions • Mathematical Expectation, Generation Functions and Law of Large Numbers • Theoretical Discrete Distributions • Theoretical Continuous Distributions • Curve Fitting and Principle of Least Squares • Correlation and Regression • Theory of Attributes • Sampling and Large Sample Tests • Chisquare Distribution • Exact, Sampling Distribution • Theory of Estimation • Testing of Hypothesis • Analysis of Variance • Design of Experiments • Design of Sample Surveys • Tables.

Fundamentals of Applied Statistics

S.C. GUPTA V.K. KAPOOR

Third Edition 2001	Pages *xvi* + 628	ISBN 81-7014-151-6	**Rs 110.00**

Special Features

- The book provides comprehensive and exhaustive theoretical discussion.
- All basic concepts have been explained in an easy and understandable manner.
- 125 stimulating problems selected from various university examinations have been solved.
- It conforms to the latest syllabi of B.Sc. (Hons.) and post-graduate examination in Statistics, Agriculture and Economics.

Contents

Statistical Quality Control • Analysis of Time Series (Mathematical Treatment) • Index Number • Demand Analysis, Price and Income Elasticity • Analysis of Variance
Design of Experiments, Completely Randomised Design • Randomised Block Design, Latin Square Design • Factorial Designs and Confounding.
Design of Sample Surveys (Mathematical Treatment) • Sample Random Sampling, Stratified Sampling, Systematic Sampling, Multi-stage Sampling • Educational and Psychological Statistics • Vital Statictical Methods.

CHAPTER ONE

Introduction – Meaning and Scope

1·1. Origin and Development of Statistics. Statistics, in a sense, is as old as the human society itself. Its origin can be traced to the old days when it was regarded as the 'science of State-craft' and was the by-product of the administrative activity of the State. The word 'Statistics' seems to have been derived from the Latin word '*status*' or the Italian word '*statista*' or the German word '*statistik*' each of which means a 'political state'. In ancient times, the government used to collect the information regarding the population and 'property or wealth' of the country – the former enabling the government to have an idea of the manpower of the country (to safeguard itself against external aggression, if any), and the latter providing it a basis for introducing news taxes and levies.

In India, an efficient system of collecting official and administrative statistics existed even more than 2,000 years ago, in particular, during the reign of Chandra Gupta Maurya (324 - 300 B.C.). From Kautilya's *Arthshastra* it is known that even before 300 B.C. a very good system of collecting 'Vital Statistics' and registration of births and deaths was in vogue. During Akbar's reign (1556 - 1605 A.D.), Raja Todarmal, the then land and revenue minister, maintained good records of land and agricultural statistics. In *Aina-e-Akbari* written by Abul Fazl (in 1596 - 97), one of the nine gems of Akbar, we find detailed accounts of the administrative and statistical surveys conducted during Akbar's reign.

In Germany, the systematic collection of official statistics originated towards the end of the 18th century when, in order to have an idea of the relative strength of different German States, information regarding population and output – industrial and agricultural – was collected. In England, statistics were the outcome of Napoleonic Wars. The Wars necessitated the systematic collection of numerical data to enable the government to assess the revenues and expenditure with greater precision and then to levy new taxes in order to meet the cost of war.

Seventeenth century saw the origin of the 'Vital Statistics.' Captain John Grant of London (1620 - 1674), known as the 'father' of Vital Statistics, was the first man to study the statistics of births and deaths. Computation of mortality tables and the calculation of expectation of life at different ages by a number of persons, *viz.*, Casper Newman, Sir William Petty (1623 - 1687), James Dodson, Dr. Price, to mention only a few, led to the idea of 'life insurance' and the first life insurance institution was founded in London in 1698.

The theoretical development of the so-called modern statistics came during the mid-seventeenth century with the introduction of 'Theory of Probability' and 'Theory of Games and Chance', the chief contributors being mathematicians and gamblers of France, Germany and England. The French mathematician Pascal (1623 - 1662), after lengthy correspondence with another French mathematician

P. Fermat (1601 - 1665) solved the famous 'Problem of Points' posed by the gambler Chevalier de - Mere. His study of the problem laid the foundation of the theory of probability which is the backbone of the modern theory of statistics. Pascal also investigated the properties of the co-efficients of binomial expansions and also invented mechanical computation machine. Other notable contributors in this field are : James Bernouli (1654 - 1705), who wrote the first treatise on the 'Theory of Probability'; De-Moivre (1667 - 1754) who also worked on probabilities and annuities and published his important work "The Doctrine of Chances" in 1718, Laplace (1749 - 1827) who published in 1782 his monumental work on the theory of probability, and Gauss (1777 - 1855), perhaps the most original of all writers on statistical subjects, who gave the principle of least squares and the normal law of errors. Later on, most of the prominent mathematicians of 18th, 19th and 20th centuries, *viz.*, Euler, Lagrange, Bayes, A. Markoff, Khintchin, Kolmogoroff, to mention only a few, added to the contributions in the field of probability.

Modern veterans in the development of the subject are Englishmen. Francis Galton (1822-1921), with his works on 'regression', pioneered the use of statistical methods in the field of Biometry. Karl Pearson (1857-1936), the founder of the greatest statistical laboratory in England (1911), is the pioneer in correlational analysis. His discovery of the 'chi square test', the first and the most important of modern tests of significance, won for Statistics a place as a science. In 1908 the discovery of Student's '*t*' distribution by W.S. Gosset who wrote under the pseudonym of 'Student' ushered in an era of exact sample tests (small samples).

Sir Ronald A. Fisher (1890 - 1962), known as the 'Father of Statistics', placed Statistics on a very sound footing by applying it to various diversified fields, such as genetics, biometry, education, agriculture, etc. Apart from enlarging the existing theory, he is the pioneer in introducing the concepts of 'Point Estimation' (efficiency, sufficiency, principle of maximum likelihood, etc.), 'Fiducial Inference' and 'Exact Sampling Distributions.' He also pioneered the study of 'Analysis of Variance' and 'Design of Experiments.' His contributions won for Statistics a very responsible position among sciences.

1·2. Definition of Statistics. Statistics has been defined differently by different authors from time to time. The reasons for a variety of definitions are primarily two. *First*, in modern times the field of utility of Statistics has widened considerably. In ancient times Statistics was confined only to the affairs of State but now it embraces almost every sphere of human activity. Hence a number of old definitions which were confined to a very narrow field of enquiry were replaced by new definitions which are much more comprehensive and exhaustive. *Secondly*, Statistics has been defined in two ways. Some writers define it as '*statistical data*', *i.e.*, numerical statement of facts, while others define it as '*statistical methods*', *i.e.*, complete body of the principles and techniques used in collecting and analysing such data. Some of the important definitions are given below.

Statistics as 'Statistical Data'

Webster defines Statistics as *"classified facts representing the conditions of the people in a State ... especially those facts which can be stated in numbers or in any other tabular or classified arrangement."* This definition, since it confines Statistics only to the data pertaining to State, is inadequate as the domain of Statistics is much wider.

Bowley defines Statistics as *" numerical statements of facts in any department of enquiry placed in relation to each other."*

A more exhaustive definition is given by Prof. Horace Secrist as follows :

" By Statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other."

Statistics as Statistical Methods

Bowley himself defines Statistics in the following three different ways :

(i) Statistics may be called the science of counting.

(ii) Statistics may rightly be called the science of averages.

(iii) Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations.

But none of the above definitions is adequate. The *first* because statistics is not merely confined to the collection of data as other aspects like presentation, analysis and interpretation, etc., are also covered by it. The *second*, because averages are only a part of the statistical tools used in the analysis of the data, others being dispersion, skewness, kurtosis, correlation, regression, etc. The *third*, because it restricts the application of Statistics to sociology alone while in modern days Statistics is used in almost all sciences – social as well as physical.

According to Boddington, " *Statistics is the science of estimates and probabilities.*" This also is an inadequate definition since probabilities and estimates constitute only a part of the statistical methods.

Some other definitions are :

" The science of Statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates." – **King.**

" Statistics is the science which deals with collection, classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomenon." – **Lovitt.**

Perhaps the best definition seems to be one given by Croxton and Cowden, according to whom Statistics may be defined as *" the science which deals with the collection, analysis and interpretation of numerical data."*

1·3. Importance and Scope of Statistics. In modern times, Statistics is viewed not as a mere device for collecting numerical data but as a means of developing sound techniques for their handling and analysis and drawing valid inferences from them. As such it is not confined to the affairs of the State but is intruding constantly into various diversified spheres of life – social, economic and political. It is now finding wide applications in almost all sciences – social as well as physical – such as biology, psychology, education, economics, business management, etc. It is hardly possible to enumerate even a single department of human activity where statistics does not creep in. It has rather become indispensable in all phases of human endeavour.

Statistics and Planning. Statistics is indispensable to planning. In the modern age which is termed as 'the age of planning', almost all over the world, goernments, particularly of the budding economies, are resorting to planning for the economic development. In order that planning is successful, it must be based soundly on the correct analysis of complex statistical data.

Statistics and Economics. Statistical data and technique of statistical analysis have proved immensely useful in solving a variety of economic problems, such as wages, prices, analysis of time series and demand analysis. It has also facilitated the development of economic theory. Wide applications of mathematics and statistics in the study of economics have led to the development of new disciplines called Economic Statistics and Econometrics.

Statistics and Business. Statistics is an indispensable tool of production control also. Business executives are relying more and more on statistical techniques for studying the needs and the desires of the consumers and for many other purposes. The success of a businessman more or less depends upon the accuracy and precision of his statistical forecasting. Wrong expectations, which may be the result of faulty and inaccurate analysis of various causes affecting a particular phenomenon, might lead to his disaster. Suppose a businessman wants to manufacture readymade garments. Before starting with the production process he must have an overall idea as to 'how many garments are to be manufactured', 'how much raw material and labour is needed for that', and 'what is the quality, shape, colour, size, etc., of the garments to be manufactured'. Thus the formulation of a production plan in advance is a must which cannot be done without having quantitative facts about the details mentioned above. As such most of the large industrial and commercial enterprises are employing trained and efficient statisticians.

Statistics and Industry. In industry, Statistics is very widely used in 'Quality Control'. In production engineering, to find whether the product is conforming to specifications or not, statistical tools, *viz.*, inspection plans, control charts, etc., are of extreme importance. In inspection plans we have to resort to some kind of sampling – a very important aspect of Statistics.

Statistics and Mathematics. Statistics and mathematics are very intimately related. Recent advancements in statistical techniques are the outcome of wide applications of advanced mathematics. Main contributors to statistics, namely, Bernouli, Pascal, Laplace, De-Moivre, Gauss, R. A. Fisher, to mention only a few, were primarily talented and skilled mathematicians. Statistics may be regarded as that branch of mathematics which provided us with systematic methods of analysing a large number of related numerical facts. According to Connor, " Statistics is a branch of Applied Mathematics which specialises in data." Increasing role of mathematics in statistical analysis has resulted in a new branch of Statistics called Mathematical Statistics.

Statistics and Biology, Astronomy and Medical Science. The association between statistical methods and biological theories was first studied by Francis Galton in his work in 'Regression'. According to Prof. Karl Pearson, the whole 'theory of heredity' rests on statistical basis. He says, " *The whole problem of evolution is a problem of vital statistics, a problem of longevity, of fertility, of health, of disease and it is impossible for the Registrar General to discuss the national mortality without an enumeration of the population, a classification of deaths and knowledge of statistical theory."*

In **astronomy**, the theory of Gaussian 'Normal Law of Errors' for the study of the movement of stars and planets is developed by using the 'Principle of Least Squares'.

In **medical science** also, the statistical tools for the collection, presentation and analysis of observed facts relating to the causes and incidence of diseases and the results obtained from the use of various drugs and medicines, are of great importance. Moreover, the efficacy of a manufacutured drug or injection or medicine is tested by using the 'tests of significance' – (t-test).

Statistics and Psychology and Education. In education and psychology, too, Statistics has found wide applications, *e.g.*, to determine the reliability and validity of a test, 'Factor Analysis', etc., so much so that a new subject called 'Psychometry' has come into existence.

Statistics and War. In war, the theory of 'Decision Functions' can be of great assistance to military and technical personnel to plan 'maximum destruction with minimum effort'.

Thus, we see that the science of Statistics is associated with almost all the sciences – social as well as physical. Bowley has rightly said, " *A knowledge of Statistics is like a knowledge of foreign language or of algebra; it may prove of use at any time under any circumstance."*

1·4. Limitations of Statistics. Statistics, with its wide applications in almost every sphere of human activity; is not without limitations. The following are some of its important limitations :

(i) Statistics is not suited to the study of qualitative phenomenon. Statistics, being a science dealing with a set of numerical data, is applicable to the study of only those subjects of enquiry which are capable of quantitative measurement. As such, qualitative phenomena like honesty, poverty, culture, etc., which cannot be expressed numerically, are not capable of direct statistical analysis. However, statistical techniques may be applied indirectly by first reducing the qualitative expressions to precise quantitative terms. For example, the intelligence of a group of candidates can be studied on the basis of their scores in a certain test.

(ii) Statistics does not study individuals. Statistics deals with an aggregate of objects and does not give any specific recognition to the individual items of a series. Individual items, taken separately, do not constitute statistical data and are meaningless for any statistical enquiry. For example, the individual figures of agricultural production, industrial output or national income of any country for a particular year are meaningless unless, to facilitate comparison, similar figures of other countries or of the same country for different years are given. Hence, statistical analysis is suited to only those problems where group characteristics are to be studied.

(iii) Statistical laws are not exact. Unlike the laws of physical and natural sciences, statistical laws are only approximations and not exact. On the basis of statistical analysis we can talk only in terms of probability and chance and not in terms of certainty. Statistical conclusions are not universally true – they are true only on an average. For example, let us consider the statement :" It has been found that 20 % of a certain surgical operations by a particular doctor are successful." The statement does not imply that if the doctor is to operate on 5 persons on any day and four of the operations have proved fatal, the fifth must be a success. It may happen that fifth man also dies of the operation or it may also happen that of the five operations on any day, 2 or 3 or even more may be successful. By the statement we mean that as number of operations becomes larger and larger we should expect, on the average, 20 % operations to be successful.

(iv) Statistics is liable to be misused. Perhaps the most important limitation of Statistics is that it must be used by experts. As the saying goes, " Statistical methods are the most dangerous tools in the hands of the inexperts. Statistics is one of those sciences whose adepts must exercise the self-restraint of an artist." The use of statistical tools by inexperienced and untrained persons might lead to very fallacious conclusions. One of the greatest shortcomings of Statistics is that they do not bear on their face the label of their quality and as such can be moulded and manipulated in any manner to support one's way of argument and reasoning. As King says, *" Statistics are like clay of which one can make a god or devil as one pleases."* The requirement of experience and skill for judicious use of statistical methods restricts their use to experts only and limits the chances of the mass popularity of this useful and important science.

1·5. Distrust of Statistics. We often hear the following interesting comments on Statistics :

(i) 'An ounce of truth will produce tons of Statistics',

(ii) 'Statistics can prove anything',

(iii) 'Figures do not lie. Liars figure',

(iv) 'If figures say so it can't be otherwise',

(v) 'There are three type of lies – lies, demand lies, and Statistics – wicked in the order of their naming, and so on.

Some of the reasons for the existence of such divergent views regarding the nature and function of Statistics are as follows :

(i) Figures are innocent, easily believable and more convincing. The facts supported by figures are psychologically more appealing.

(ii) Figures put forward for arguments may be inaccurate or incomplete and thus might lead to wrong inferences.

(iii) Figures, though accurate, might be moulded and manipulated by selfish persons to conceal the truth and present a distorted picture of facts to the public to meet their selfish motives. When the skilled talkers, writers or politicians through their forceful writings and speeches or the business and commercial enterprises through advertisements in the press mislead the public or belie their expectations by quoting wrong statistical statements or manipulating statistsical data for personal motives, the public loses its faith and belief in the science of Statistics and starts condemning it. We cannot blame the layman for his distrust of Statistics, as he, unlike statistician, is not in a position to distinguish between valid and invalid conclusions from statistical statements and analysis.

It may be pointed out that Statistics neither proves anything nor disproves anything. It is only a tool which if rightly used may prove extremely useful and if misused, might be disastrous. According to Bowley, *"Statistics only furnishes a tool, necessary though imperfect, which is dangerous in the hands of those who do not know its use and its deficiencies."* It is not the subject of Statistics that is to be blamed but those people who twist the numerical data and misuse them either due to ignorance or deliberately for personal selfish motives. As King points out, *"Science of Statistics is the most useful servant but only of great value to those who understand its proper use."*

We discuss below a few interesting examples of misrepresentation of statistical data.

(i) A statistical report : "The number of accidents taking place in the middle of the road is much less than the number of accidents taking place on its side. Hence it is safer to walk in the middle of the road." This conclusion is obviously wrong since we are not given the proportion of the number of accidents to the number of persons walking in the two cases.

(ii) "The number of students taking up Mathematics Honours in a University has increased 5 times during the last 3 years. Thus, Mathematics is gaining popularity among the students of the university." Again, the conclusion is faulty since we are not given any such details about the other subjects and hence comparative study is not possible.

(iii) "99% of the people who drink alcohol die before attaining the age of 100 years. Hence drinking is harmful for longevity of life." This statement, too, is incorrect since nothing is mentioned about the number of persons who do not drink alcohol and die before attaining the age of 100 years.

Thus, statistical arguments based on incomplete data often lead to fallacious conclusions.

FREQUENCY DISTRIBUTIONS AND MEASURES OF CENTRAL TENDENCY

2·1. Frequency Distributions. When observations, discrete or continuous, are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. Let us consider the marks in Statistics obtained by 250 candidates selected at random from among those appearing in a certain examination.

TABLE 1 : MARKS IN STATISTICS OF 250 CANDIDATES

32	47	41	51	41	30	39	18	48	53
54	32	31	46	15	37	32	56	42	48
38	26	50	40	38	42	35	22	62	51
44	21	45	31	37	41	44	18	37	47
68	41	30	52	52	60	42	38	38	34
41	53	48	21	28	49	42	36	41	29
30	33	37	35	29	37	38	40	32	49
43	32	24	38	38	22	41	50	17	46
46	50	26	15	23	42	25	52	38	46
41	38	40	37	40	48	45	30	28	31
40	33	42	36	51	42	56	44	35	38
31	51	45	41	50	53	50	32	45	48
40	43	40	34	34	44	38	58	49	28
40	45	19	24	34	47	37	33	37	36
36	32	61	30	44	43	50	31	38	45
46	40	32	34	44	54	35	39	31	48
48	50	43	55	43	39	41	48	53	34
32	31	42	34	34	32	33	24	43	39
40	50	27	47	34	44	34	33	47	42
17	42	57	35	38	17	33	46	36	23
48	50	31	58	33	44	26	29	31	37
47	55	57	37	41	54	42	45	47	43
37	52	47	46	44	50	44	38	42	19
52	45	23	41	47	33	42	24	48	39
48	44	60	38	38	44	38	43	40	48

This representation of the data does not furnish any useful information and is rather confusing to mind. A better way may be to express the figures in an ascending or descending order of magnitude, commonly termed as *array*. But this does not reduce the bulk of the data. A much better representation is given on the next page.

A bar (|) called *tally mark* is put against the number when it occurs. Having occurred four times, the fifth occurrence is represented by putting a cross tally (/) on the first four tallies. This technique facilitates the counting of the tally marks at the end.

The representation of the data as above is known as *frequency distribution.* Marks are called the *variable* (*x*) and the *'number of students'* against the marks is known as the *frequency* (*f*) of the variable. The word *'frequency'* is derived from 'how frequently' a variable occurs. For example, in the above case the frequency of 31 is 10 as there are ten students getting 31 marks. This representation, though better than an array', does not condense the data much and it is quite cumbersome to go through this huge mass of data.

TABLE 2

Marks	*No. of Students Tally Marks*	*Total Frequency*	*Marks*	*No. of Students Tally Marks*	*Total Frequency*
15	II	= 2	40	卌 卌 I	= 11
17	III	= 3	41	卌 卌	= 10
18	II	= 2	42	卌 卌 III	= 13
19	II	= 2	43	卌 III	= 8
21	II	= 2	44	卌 卌 II	= 12
22	II	= 2	45	卌 II	= 7
23	III	= 3	46	卌 II	= 7
24	IIII	= 4	47	卌 III	= 8
25	I	= 1	48	卌 卌 II	= 12
26	III	= 3	49	III	= 3
27	I	= 1	50	卌 卌	= 10
28	III	= 3	51	IIII	= 4
29	II	= 2	52	卌	= 5
30	卌	= 5	53	IIII	= 4
31	卌 卌	= 10	54	III	= 3
32	卌 卌	= 10	55	II	= 2
33	卌 III	= 8	56	II	= 2
34	卌 卌 I	= 11	57	II	= 2
35	卌	= 5	58	II	= 2
36	卌	= 5	60	III	= 3
37	卌 卌 II	= 12	61	I	= 1
38	卌 卌 卌 II	= 17	62	I	= 1
39	卌 I	= 6	68	I	= 1

If the identity of the individuals about whom a particular information is taken is not relevant, nor the order in which the observations arise, then the first real step of condensation is to divide the observed range of variable into a suitable number of *class-intervals* and to record the number of observations in each class. For example, in the above case, the data may be expressed as shown in Table 3.

Such a table showing the distribution of the frequencies in the different classes is called a *frequency table* and the manner in which the class frequencies are distributed over the class intervals is called the *grouped frequency distribution* of the variable.

Remark. The classes of the type 15—19, 20—24, 25—29 etc., in which both the upper and lower limits are included are called *'inclusive classes'*. For example the class 20—24, includes

TABLE 3 : FREQUENCY TABLE

Marks (*x*)	*No. of students.* (*f*)
15 — 19	9
20 — 24	11
25 — 29	10
30 — 34	44
35 — 39	45
40 — 44	54
45 — 49	37
50 — 54	26
55 — 59	8
60 — 64	5
65 — 69	1
Total	250

all the values from 20 to 24, both inclusive and the classification is termed as *inclusive type classification.*

In spite of great importance of classification in statistical analysis, no hard and fast rules can be laid down for it. The following points may be kept in mind for classification :

(i) The classes should be clearly defined and should not lead to any ambiguity.

(ii) The classes should be exhaustive, *i.e.*, each of the given values should be included in one of the classes.

(iii) The classes should be mutually exclusive and non-overlapping.

(iv) The classes should be of equal width. The principle, however, cannot be rigidly followed. If the classes are of varying width, the different class frequencies will not be comparable. Comparable figures can be obtained by dividing the value of the frequencies by the corresponding widths of the class intervals. The ratios thus obtained are called *'frequency densities'*.

(v) Indeterminate classes, *e.g.*, the open-end classes, less than '*a*' or greater than '*b*' should be avoided as far as possible since they create difficulty in analysis and interpretation.

(vi) The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15. However, the number of classes may be more than 15 depending upon the total frequency and the details required, but it is desirable that it is not less than 5 since in that case the classification may not reveal the essential characteristics of the population. The following formula due to Struges may be used to determine an approximate number k of classes :

$$k = 1 + 3{\cdot}322 \log_{10} N,$$

where N is the total frequency.

The Magnitude of the Class Interval

Having fixed the number of classes, divide the range (the difference between the greatest and the smallest observation) by it and the nearest integer to this value gives the magnitude of the class interval. Broad class intervals (*i.e.*, less number of classes) will yield only rough estimates while for high degree of accuracy small class intervals (*i.e.*, large number of classes) are desirable.

Class Limits

The class limits should be chosen in such a way that the mid-value of the class interval and actual average of the observations in that class interval are as near to each other as possible. If this is not the case then the classification gives a distorted picture of the characteristics of the data. If possible, class limits should be located at the points which are multiple of 0, 2, 5, 10,... etc., so that the midpoints of the classes are the common figures, *viz.*, 0, 2, 5, 10..., etc., the figures capable of easy and simple analysis.

2·1·1. Continuous Frequency Distribution. If we deal with a continuous variable, it is not possible to arrange the data in the class intervals of above type. Let us consider the distribution of age in years. If class intervals are 15—19, 20—24 then the persons with ages between 19 and 20 years are not taken into consideration. In such a case we form the class intervals as shown below.

Age in years
Below 5
5 or more but less than 10
10 or more but less than 15
15 or more but less than 20
20 or more but less than 25
and so on.

Here all the persons with any fraction of age are included in one group or the other. For practical purpose we re-write the above classes as

0 — 5
5 — 10
10 — 15
15 — 20
20 — 25

This form of frequency distribution is known as *continuous frequency distribution.*

It should be clearly understood that in the above classes, the upper limits of each class are excluded from the respective classes. Such classes in which the upper limits are excluded from the respective classes and are included in the immediate next class are known as *'exclusive classes'* and the classification is termed as *'exclusive type classification'*.

2·2. Graphic Representation of a Frequency Distribution. It is often useful to represent a frequency distribution by means of a diagram which makes the unwieldy data intelligible and conveys to the eye the general run of the observations. Diagrammatic representation also facilitates the comparison of two or more frequency distributions. We consider below some important types of graphic representation.

2·2·1. Histogram. In drawing the histogram of a given continuous frequency distribution we first mark off along the x–axis all the class intervals on a suitable scale. On each class interval erect rectangles with heights proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class. If, however, the classes are of unequal width then the height of the rectangle wili be proportional to the ratio of the frequencies to the width of the classes. The diagram of continuous rectangles so obtained is called *histrogram*.

Remarks. 1. To draw the histogram for an ungrouped frequency distribution of a variable we shall have to assume that the frequency corresponding to the variate value x is spread over the interval $x - h/2$ to $x + h/2$, where h is the jump from one value to the next.

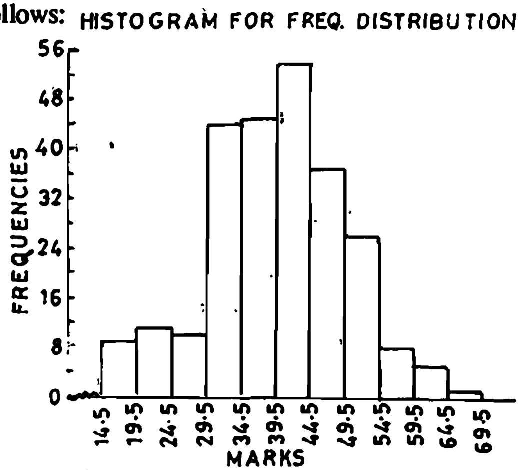
2. If the grouped frequency distribution is not continuous, first it is to be converted into continuous distribution and then the histrogam is drawn.

3. Although the height of each rectangle is proportional to the frequency of the corresponding class, the height of a fraction of the rectangle is not proportional to the frequency of the corresponding fraction of the class, so that histogram cannot be directly used to read frequency over a fraction of a class interval.

4. The histogram of the distribution of marks of 250 students in *Table* 3 (page 2·2) is obtained as follows.

Since the grouped frequency distribution is not continuous, we first convert it into a continuous distribution as follows:

Marks	*No. of Students*
14·5–19·5	9
19·5–24·5	11
24·5–29·5	10
29·5–34·5	44
34·5–39·5	45
39·5–44·5	54
44·5–49·5	37
49·5–54·5	26
54·5–59·5	8
59·5–64·5	5
64·5–69·5	1



Remark. The upper and lower class limits of the *new exclusive type classes* are known as *class boundaries.*

If d is the gap between the upper limit of any class and the lower limit of the succeeding class, the class boundaries for any class are then given by :

$$\text{Upper class boundary} = \text{Upper class limit} + \frac{d}{2}$$

$$\text{Lower class boundary} = \text{Lower class limit} - \frac{d}{2}$$

2·2·2. **Frequency Polygon.** For an ungrouped distribution, the frequency polygon is obtained by plotting points with abscissa as the variate values and the ordinate as the corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the abscissa of points are mid-values of the class intervals. For equal class intervals the frequency polygon can be obtained by joining the middle points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width the polygon can be approximated by a smooth curve. The frequency curve can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

2·3. Averages or Measures of Central Tendency or Measures of Location. According to Professor Bowley, averages are "statistical constants which enable us to comprehend in a single effort the significance of the whole." They give us an idea about the concentration of the values in the central part of the distribution. Plainly speaking, an average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of central tendency that are in common use:

(i) Arithmetic Mean or simply Mean , (ii) Median,
(iii) Mode, (iv) Geometric Mean, and (v) Harmonic Mean.

2·4. Requisites for an Ideal Measure of Central Tendency. According to Professor Yule, the following are the characteristics to be satisfied by an ideal measure of central tendency :

(i) It should be rigidly defined.

(ii) It should be readily comprehensible and easy to calculate.

(iii) It should be based on all the observations.

(iv) It should be suitable for further mathematical treatment. By this we mean that if we are given the averages and sizes of a number of series, we should be able to calculate the average of the composite series obtained on combining the given series.

(v) It should be affected as little as possible by fluctuations of sampling.

In addition to the above criteria, we may add the following (which is not due to Prof.Yule) :

(vi) It should not be affected much by extreme values.

2·5. Arithmetic Mean. Arithmetic mean of a set of observastions is their sum divided by the number of observations, *e.g.*, the arithmetic mean $\bar{x}$ of n observations x_1 , x_2 , ... , x_n is given by

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n \right) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

In case of frequency distribution $x_i \mid f_i$, $i = 1, 2, ..., n$, where f_i is the frequency of the variable x_i ;

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots f_n} = \frac{\sum\limits_{i=1}^{n} f_i x_i}{\sum\limits_{i=1}^{n} f_i} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i, \quad \left[\sum_{i=1}^{n} f_i = N \right] \qquad \ldots(2 \cdot 1)$$

In case of grouped or continuous frequency distribution, x is taken as the mid-value of the corresponding class.

Remark. The symbol Σ is the letter capital sigma of the Greek alphabet and is used in mathematics to denote the sum of values.

Example 2·1. *(a) Find the arithmetic mean of the following frequency distribution:*

x :	1	2	3	4	5	6	7
f :	5	9	12	17	14	10	6

(b) Calculate the arithmetic mean of the marks from the following table :

Marks :	0–10	10–20	20–30	30–40	40–50	50–60
No. of students :	12	18	27	20	17	6

Solution. *(a)*

x	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
	73	299

$$\therefore \qquad \bar{x} = \frac{1}{N} \Sigma fx = \frac{299}{73} = 4{\cdot}09$$

(b)

Marks	*No. of students* (f)	*Mid - point* (x)	fx
0–10	12	5	60
10–20	18	15	270
20–30	27	25	675
30–40	20	35	700
40–50	17	45	765
50–60	6	55	330
Total	100		2,800

$$\text{Arithmetic mean or } \bar{x} = \frac{1}{N} \Sigma fx = \frac{1}{100} \times 2{,}800 = 28$$

It may be noted that if the values of x or (and) f are large, the calculation of mean by formula *(2·1)* is quite time-consuming and tedious. The arithmetic is reduced to a great extent by taking the deviations of the given values from any arbitrary point 'A', as explained below.

Let $d_i = x_i - A$, then $f_i d_i = f_i (x_i - A) = f_i x_i - Af_i$

Summing both sides over i from 1 to n, we get

$$\sum_{i=1}^{n} f_i d_i = \sum_{i=1}^{n} f_i x_i - A \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i x_i - A \,.\, N.$$

$$\Rightarrow \qquad \frac{1}{N}\sum_{i=1}^{n} f_i d_i = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - A = \bar{x} - A,$$

where $\bar{x}$ is the arithmetic mean of the distribution.

$$\therefore \qquad \bar{x} = A + \frac{1}{N}\sum_{i=1}^{n} f_i d_i \qquad ...(2{\cdot}2)$$

This formula is much more convenient to apply than formula (2·1).

Any number can serve the purpose of arbitrary point 'A' but, usually, the value of x corresponding to the middle part of the distribution will be much more convenient.

In case of grouped or continuous frequency distribution, the arithmetic is reduced to a still greater extent by taking

$$d_i = \frac{x_i - A}{h},$$

where A is an arbitrary point and h is the common magnitude of class interval. In this case, we have

$$h\, d_i = x_i - A,$$

and proceeding exactly similarly as above, we get

$$\bar{x} = A + \frac{h}{N}\sum_{i=1}^{n} f_i d_i \qquad ...(2{\cdot}3)$$

Example 2·2. *Calculate the mean for the following frequency distribution.*

Class-interval :	*0–8*	*8–16*	*16–24*	*24–32*	*32–40*	*40–48*
Frequency :	*8*	*7*	*16*	*24*	*15*	*7*

Solution.

Class–interval	*mid–value* (x)	*Frequency* (f)	$d = (x-A)/h$	fd
0–8	4	8	–3	–24
8–16	12	7	–2	–14
16–24	20	16	–1	–16
24–32	28	24	0	0
32–40	36	15	1	15
40–48	44	7	2	14
		77		–25

Here we take $A = 28$ and $h = 8$.

$$\therefore \qquad \bar{x} = A + \frac{h\,\Sigma f d}{N} = 28 + \frac{8 \times (-25)}{77} = 28 - \frac{200}{77} = 25{\cdot}404$$

2·5·1. Properties of Arithmetic Mean

Property 1. *Algebraic sum of the deviations of a set of values from their arithmetic mean is zero. If* $x_i \mid f_i$, $i = 1, 2, ..., n$ *is the frequency distribution, then*

$$\sum_{i=1}^{n} f_i\,(x_i - \bar{x}) = 0, \; \bar{x} \textit{ being the mean of distribution.}$$

Proof. $\sum_i f_i (x_i - \bar{x}) = \sum_i f_i\, x_i - \bar{x} \sum f_i = \sum_i f_i\, x_i - \bar{x}\,.N$

Also $\bar{x} = \dfrac{\sum_i f_i x_i}{N} \Rightarrow \sum_i f_i x_i = N\,\bar{x}$

Hence $\sum_{i=1}^{n} f_i (x_i - \bar{x}) = N\,.\bar{x} - \bar{x}\,.N = 0$

Property 2. *The sum of the squares of the deviations of a set of values is minimum when taken about mean.*

Proof. For the frequency distribution $x_i \mid f_i$, $i = 1, 2, ..., n$, let

$$Z = \sum_{i=1}^{n} f_i (x_i - A)^2,$$

be the sum of the squares of the deviations of given values from any arbitrary point 'A'. We have to prove that Z is minimum when $A = \bar{x}$.

Applying the principle of maxima and minima from differential calculus, Z will be minimum for variations in A if

$$\frac{\partial Z}{\partial A} = 0 \text{ and } \frac{\partial^2 Z}{\partial A^2} > 0$$

Now $\dfrac{\partial Z}{\partial A} = -\,2 \sum_i f_i (x_i - A) = 0 \Rightarrow \sum_i f_i (x_i - A) = 0$

$$\Rightarrow \quad \sum f_i x_i - A \sum f_i = 0 \quad or \quad A = \frac{\sum f_i x_i}{N} = \bar{x}$$

Again $\dfrac{\partial^2 Z}{\partial A^2} = -\,2 \sum_i f_i (-1) = 2 \sum_i f_i = 2N > 0$

Hence Z is minimum at the point $A = \bar{x}$. This establishes the result.

Property 3. *(Mean of the composite series). If $\bar{x}_i$, (i = 1, 2, ..., k) are the means of k-component series of sizes n_i, (i = 1, 2, ..., k) respectively, then the mean $\bar{x}$ of the composite series obtained on combining the component series is given by the formula:*

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \;\dots + n_k \bar{x}_k}{n_1 + n_2 + \;\dots + n_k} = \sum_i n_i \bar{x}_i \Big/ \sum_i n_i \qquad ...(2\cdot4)$$

Proof. Let $x_{11}, x_{12}, .. x_{1n_1}$ be n_1 members of the first series ; $x_{21}, x_{22}, ...$ x_{2n_2} be n_2 members of the second series, $x_{k1}, x_{k2}, ..., x_{kn_k}$ be n_k members of the kth series. Then, by def.,

$$\left.\begin{aligned} \bar{x}_1 &= \frac{1}{n_1}(x_{11} + x_{12} + \dots + x_{1n_1}) \\ \bar{x}_2 &= \frac{1}{n_2}(x_{21} + x_{22} + \dots + x_{2n_2}) \\ &\vdots \qquad \vdots \qquad \vdots \\ \bar{x}_k &= \frac{1}{n_k}(x_{k1} + x_{k2} + \dots + x_{kn_k}) \end{aligned}\right\} \quad \dots(*)$$

The mean $\bar{x}$ of composite series of size $n_1 + n_2 + \dots + n_k$ is given by

$$\bar{x} = \frac{(x_{11} + x_{12} + \dots + x_{1n_1}) + (x_{21} + x_{22} + \dots + x_{2n_2}) + \dots + (x_{k1} + x_{k2} + \dots + x_{kn_k})}{n_1 + n_2 + \dots + n_k}$$

$$= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n_1 + n_2 + \dots n_k}, \qquad \text{[From (*)]}$$

Thus, $\bar{x} = \sum_i n_i \bar{x}_i \, / \, (\sum_i n_i)$

Example 2·3. *The average salary of male employees in a firm was Rs.520 and that of females was Rs.420. The mean salary of all the employees was Rs.500. Find the percentage of male and female employees.*

Solution. Let n_1 and n_2 denote respectively the number of male and female employees in the concern and $\bar{x}_1$ and $\bar{x}_2$ denote respectively their average salary (in rupees). Let $\bar{x}$ denote the avarage salary of all the workers in the firm.

We are given that :

$$\bar{x}_1 = 520, \quad \bar{x}_2 = 420 \quad \text{and} \quad \bar{x} = 500$$

Also we know

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\Rightarrow \quad 500(n_1 + n_2) = 520\, n_1 + 420\, n_2$$

$$\Rightarrow \quad (520 - 500)\, n_1 = (500 - 420)\, n_2$$

$$\Rightarrow \quad 20\, n_1 = 80\, n_2$$

$$\Rightarrow \quad \frac{n_1}{n_2} = \frac{4}{1}$$

Hence the percentage of male employees in the firm

$$= \frac{4}{4+1} \times 100 = 80$$

and percentage of female employees in the firm

$$= \frac{1}{4+1} \times 100 = 20$$

2·5·2. Merits and Demerits of Arithmetic Mean

Merits. *(i)* It is rigidly defined .

(ii) It is easy to understand and easy to calculate.

(iii) It is based upon all the observations.

(iv) It is amenable to algebraic treatment. The mean of the composite series in terms of the means and sizes of the component series is given by

$$\bar{x} = \sum_{i=1}^{k} n_i \bar{x}_i \Big/ \left(\sum_{i}^{k} n_i \right)$$

(v) Of all the averages, arithmetic mean is affected least by fluctuations of sampling. This property is sometimes described by saying that arithmetic mean is a *stable* average.

Thus, we see that arithmetic mean satisfies all the properties laid down by Prof. Yule for an ideal average.

Demerits. *(i)* It cannot be determined by inspection nor it can be located graphically.

(ii) Arithmetic mean cannot be used if we are dealing with qualitative characteristics which cannot be measured quantitively; such as, intelligence, honesty, beauty, etc. In such cases median (discussed later) is the only average to be used.

(iii) Arithmetic mean cannot be obtained if a single observation is missing or lost or is illegible unless we drop it out and compute the arithmetic mean of the remaining values.

(iv) Arithmetic mean is affected very much by extreme values. In case of extreme items, arithmetic mean gives a distorted picture of the distribution and no longer remains representative of the distribution.

(v) Arithmetic mean may lead to wrong conclusions if the details of the data from which it is computed are not given. Let us consider the following marks obtained by two students *A* and *B* in three tests, *viz.*, terminal test, half-yearly examination and annual examination respectively.

Marks in : →	*I Test*	*II Test*	*III Test*	*Average marks*
A	50%	60%	70%	60%
B	70%	60%	50%	60%

Thus average marks obtained by each of the two students at the end of the year are 60%. If we are given the average marks alone we conclude that the level of intelligence of both the students at the end of the year is same. This is a fallacious conclusion since we find from the data that student *A* has improved consistently while student *B* has deteriorated consistently.

(vi) Arithmetic mean cannot be calculated if the extreme class is open, *e.g.*, below 10 or above 90. Morever, even if a single observation is missing mean cannot be calculated.

(vii) In extremely asymmertrical (skewed) distribution, usually arithmetic mean is not a suitable measure of location.

2·5·3. Weighted Mean. In calculating arithmetic mean we suppose that all the items in the distribution have equal importance. But in practice this may not be so. If some items in a distribution are more important than others, then this point must be borne in mind, in order that average computed is representative of the distribution. In such cases, proper weightage is to be given to various items — the weights attached to each item being proportional to the importance of the item in the distribution. For example, if we want to have an idea of the change in cost of living of a certain group of people, then the simple mean of the prices of the commodities consumed by them will not do, since all the commodities are not equally important, *e.g.*, wheat, rice and pulses are more important than cigarettes, tea, confectionery, etc.

Let w_i be the weight attached to the item x_i, $i = 1, 2, \ldots, n$. Then we define :

Weighted arithmetic mean or weighted mean $= \sum\limits_i w_i x_i / \sum\limits_i w_i$ ... (2·5)

It may be observed that the formula for weighted mean is the same as the formula for simple mean with f_i, $(i = 1, 2, \ldots, n)$, *the frequencies replaced by* w_i, $(i = 1, 2, \ldots, n)$, *the weights.*

Weighted mean gives the result equal to the simple mean if the weights assigned to each of the variate values are equal. It results in higher value than the simple mean if smaller weights are given to smaller items and larger weights to larger items. If the weights attached to larger items are smaller and those attached to smaller items are larger, then the weighted mean results in smaller value than the simple mean.

Example 2·4. *Find the simple and weighted arithmetic mean of the first n natural numbers, the weights being the corresponding numbers.*

Solution. The first *natural* numbers are 1, 2, 3, ..., n.

X	w	wX
1	1	1^2
2	2	2^2
3	3	3^2
$\vdots$	$\vdots$	$\vdots$
n	n	n^2

We know that

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$$

$$1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Simple A.M. is

$$\overline{X} = \frac{\sum X}{n} = \frac{1 + 2 + 3 + \ldots + n}{n} = \frac{n+1}{2}$$

Weighted A.M. is

$$\overline{X}_w = \frac{\sum wX}{\sum w} = \frac{1^2 + 2^2 + \ldots + n^2}{1 + 2 + \ldots + n}$$

$$= \frac{n(n+1)(2n+1)}{6} \cdot \frac{2}{n(n+1)}$$

2·6. Median. Median of a distribution is the value of the variable which divides it into two equal parts. It is the value which exceeds and is exceeded by the same number of observations, *i.e.*, it is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a *positional* average.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms. For example, the median of the values 25, 20, 15, 35, 18, *i.e.*, 15, 18, 20, 25, 35 is 20 and the median of 8, 20, 50, 25, 15, 30, *i.e.*, of 8, 15, 20, 25, 30, 50 is $\frac{1}{2}(20+25)=22\cdot5$.

Remark. In case of even number of observations, in fact any value lying between the two middle values can be taken as median but conventionally we take it to be the mean of the middle terms.

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

(i) Find $N/2$, where $N=\sum_i f_i$.

(ii) See the (less than) cumulative frequency (*c.f.*) just greater than $N/2$.

(iii) The corresponding value of x is median.

Example 2·5. *Obtain the median for the following frequency distribution:*

x :	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
f :	*8*	*10*	*11*	*16*	*20*	*25*	*15*	*9*	*6*

Solution.

x	f	$c.f.$
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	120	

Hence $N=120 \Rightarrow N/2=60$

Cumulative frequency (*c.f.*) just greater than $N/2$, is 65 and the value of x corresponding to 65 is 5. Therefore, median is 5.

In the case of continuous frequency distribution, the class corresponding to the *c.f.* just greater than $N/2$ is called the *median class* and the value of median is obtained by the following formula :

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right) \qquad \ldots(2\cdot6)$$

where l is the lower limit of the median class,
f is the frequency of the median class,
h is the magnitude of the median class,
'c' is the $c.f.$ of the class *preceding* the median class,
and $N = \Sigma f$.

Derivation of the Median Formula (2·6). Let us consider the following continuous frequency distribution, ($x_1 < x_2 < \ldots < x_{n+1}$) :

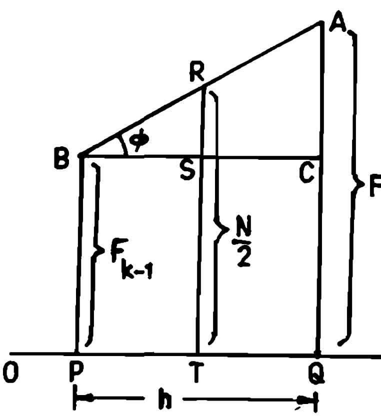
Class interval : $x_1 - x_2$, $x_2 - x_3$, ... $x_k - x_{k+1}$, ... $x_n - x_{n+1}$
Frequency : f_1 f_2 .... f_k .... f_n

The cumulaltive frequency distribution is given by :

Class interval : $x_1 - x_2$, $x_2 - x_3$, ... $x_k - x_{k+1}$, ... $x_n - x_{n+1}$
Frequency : F_1 F_2 .... F_k .... F_n

where $F_i = f_1 + f_2 + \ldots\ldots + f_i$. The class $x_k - x_{k+1}$ is the median class if and only if $F_{k-1} < N/2 \le F_k$.

Now, if we assume that the variate values are uniformly distributed over the median-class which implies that the ogive is a straight line in the median-class, then we get from the Fig. 1,



$$\tan\phi = \frac{RS}{BS} = \frac{AC}{BC}$$

i.e.
$$\frac{RT - TS}{BS} = \frac{AQ - CQ}{BC}$$

or
$$\frac{RT - BP}{BS} = \frac{AQ - BP}{PQ}$$

or
$$\frac{N/2 - F_{k-1}}{BS} = \frac{F_k - F_{k-1}}{PQ} = \frac{f_k}{h}$$

where f_k is the frequency and h the magnitude of the median class.

$$\therefore \qquad BS = \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$$

Hence

$$\text{Median} = OT = OP + PT = OP + BS$$
$$= l + \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$$

which is the required formula.

Remark. The median formula (2·6) can be used only for continuous classes without any gaps, *i.e.*, for '*exclusive type*' classification. If we are given a frequency distribution in which classes are of *'inclusive type'* with gaps, then it must be converted into a continuous *'exclusive type'* frequency distribution without any gaps before applying (2·6). This will affect the value of l in (2·6). As an illustration see Example 2·7.

Example 2·6. *Find the median wage of the following distribution :*

Wages (in Rs.) :	*20—30*	*30—40*	*40—50*	*50—60*	*60—70*
No. of labourers :	*3*	*5*	*20*	*10*	*5*

[Gorakhpur Univ. B. Sc. 1989]

Solution.

Wages (in Rs.)	*No. of labourers*	*c.f.*
20—30	3	3
30—40	5	8
40—50	20	28
50—60	10	38
60—70	5	43

Here $N/2 = 43/2 = 21{\cdot}5$. Cumulative frequency just greater than 21·5 is 28 and the corresponding class is 40–50. Thus median class is 40–50. Hence using (2·6), we get

$$\text{Median} = 40 + \frac{10}{20}(21{\cdot}5 - 8) = 40 + 6{\cdot}75 = 46{\cdot}75$$

Thus median wage is Rs. 46·75.

Example 2·7. *In a factory employing 3,000 persons, 5 per cent earn less than Rs. 3 per hour, 580 earn from Rs. 3·01 to Rs. 4·50 per hour, 30 percent earn from Rs. 4·51 to Rs. 6·00 per hour, 500 earn from Rs. 6·01 to Rs. 7·50 per hour, 20 percent earn from Rs. 7·51 to Rs. 9·00 per hour, and the rest earn Rs. 9·01 or more per hour. What is the median wage?* **[Utkal Univ. B.Sc.1992]**

Solution. The given information can be expressed in tabular form as follows.

CALCULATIONS FOR MEDIAN WAGE

Earnings (in Rs.)	*Percentage of workers*	*No. of workers (f)*	*Less than c.f.*	*Class boundaries*
less than 3	5%	$\frac{5}{100} \times 3000 = 150$	150	Below 3·005
3·01—4·50	—	580	730	3·005—4·505
4·51—6·00	30 %	$\frac{30}{100} \times 3000 = 900$	1630	4·505—6·005
6·01—7·50	—	500	2130	6·005—7·505
7·51—9·00	20 %	$\frac{20}{100} \times 3000 = 600$	2730	7·505—9·005
9·01 and above		$3000 - 2730 = 270$	$3000 = N$	9·005 and above

$N/2 = 1500$. The $c.f.$ just greater than 1500 is 1630. The corresponding class 4·51–6·00, whose class boundaries are 4·505–6·005, is the median class. Using the median formula, we get:

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - C\right) = 4{\cdot}505 + \frac{1{\cdot}5}{900}(1500 - 730)$$
$$= 4{\cdot}505 + 1{\cdot}283 \approx 5{\cdot}79$$

Hence median wage is Rs. 5·79.

Example 2·8. *An incomplete frequency distribution is given as follows:*

Variable	*Frequency*	*Variable*	*Frequency*
10—20	*12*	*50—60*	*?*
20—30	*30*	*60—70*	*25*
30—40	*?*	*70—80*	*18*
40—50	*65*	*Total*	*229*

Given that the median value is 46, determine the missing frequencies using the median formula. **[Delhi Univ. B. Sc., Oct. 1992]**

Solution. Let the frequency of the class 30—40 be f_1 and that of 50—60 be f_2.

Then $f_1 + f_2 = 229 - (12 + 30 + 65 + 25 + 18) = 79$.

Since median is given to be 46, the class 40—50 is the median class.

Hence using median formula (2·6), we get

$$46 = 40 + \frac{114{\cdot}5 - (12 + 30 + f_1)}{65} \times 10$$

$$46 - 40 = \frac{72.5 - f_1}{65} \times 10 \quad \text{or} \quad 6 = \frac{72{\cdot}5 - f_1}{6{\cdot}5}$$

$$f_1 = 72{\cdot}5 - 39 = 33{\cdot}5 \approx 34$$

[Since frequency is never fractional]

$\therefore \quad f_2 = 79 - 34 = 45$ [Since $f_1 + f_2 = 79$]

2·6·1. Merits and Demerits of Median

Merits. *(i)* It is rigidly defined.

(ii) It is easily understood and is easy to calculate. In some cases it can be located merely by inspection.

(iii) It is not at all affected by extreme values.

(iv) It can be calculated for distributions with open-end classes.

Demerits. *(i)* In case of even number of observations median cannot be determined exactly. We merely estimate it by taking the mean of two middle terms.

(ii) It is not based on all the observations. For example, the median of 10, 25, 50, 60 and 65 is 50. We can replace the observations 10 and 25 by any two values which are smaller than 50 and the observations 60 and 65 by any two values greater than 50 without affecting the value of median. This property is sometimes described

by saying that median is *insensitive*.

(iii) It is not amenable to algebraic treatment.

(iv) As compared with mean, it is affected much by fluctuations of sampling.

Uses. *(i)* Median is the only average to be used while dealing with qualitative data which cannot be measured quantitatively but still can be arranged in ascending or descending order of magnitude, *e.g.*, to find the average intelligence or average honesty among a group of people.

(ii) It is to be used for determining the typical value in problems concerning wages, distribution of wealth, etc.

2·7. Mode. Let us cosider the following statements :

(i) The average height of an Indian (male) is 5′–6″.

(ii) The average size of the shoes sold in a shop is 7.

(iii) An average student in a hostel spends Rs.150 p.m.

In all the above cases, the average referred to is mode. Mode is the value which occurs most frequently in a set of observastions and around which the other items of the set cluster densely. In other words, mode is the value of the variable which is predominant in the series. Thus in the case of discrete frequency distribution mode is the value of x corresponding to maximum frequency. For example, in the following frequency distribution :

x :	1	2	3	4	5	6	7	8
f :	4	9	16	25	22	15	7	3

the value of x corresponding to the maximum frequency, *viz.*, 25 is 4. Hence mode is 4.

But in any one (or more) of the following cases :

(i) if the maximum frequency is repeated,

(ii) if the maximum frequency occurs in the very beginning or at the end of the distribution, and

(iii) if there are irregularities in the distribution,

the value of mode is determined by the *method of grouping*, which is illustrated below by an example.

Example 2·9. *Find the mode of the following frequency distribution :*

Size (x) :	1	2	3	4	5	6	7	8	9	10	11	12
Frequency (f) :	3	8	15	23	35	40	32	28	20	45	14	6

Solution. Here we see that the distribution is not regular since the frequencies are increasing steadily up to 40 and then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution. Here we cannot say that since maximum frequency is 45, mode is 10. Here we shall locate mode by the method of grouping as explained below :

The frequencies in column *(i)* are the original frequencies. Column *(ii)* is obtained by combining the frequencies two by two. If we leave the first frequency and combine the remaining frequencies two by two we get column *(iii)*. Combining

Size (x)	Frequency (i)	(ii)	(iii)	(iv)	(v)	(vi)
1	3	11		26		
2	8		23		46	
3	15	38				73
4	23		58	**98**		
5	35	**75**			**107**	
6	40		**72**			**100**
7	32	60		80		
8	28		48		93	
9	20	65				79
10	**45**		59	65		
11	14	20				
12	6					

the frequencies two by two after leaving the first two frequencies results in a repetition of column *(ii)*. Hence, we proceed to combine the frequencies three by three, thus getting column *(iv)*. The combination of frequencies three by three after leaving the first frequency results in column *(v)* and after leaving the first two frequencies results in column *(vi)*.

The maximum frequency in each column is given in black type. To find mode we form the following table :

ANALYSIS TABLE

Column Number *(1)*	*Maximum Frequency* *(2)*	*Value or combination of values of x giving max. frequency in (2)* *(3)*
(i)	45	10
(ii)	75	5, 6
(iii)	72	6, 7
(iv)	98	4, 5, 6,
(v)	107	5, 6, 7
(vi)	100	6, 7, 8

On examining the values in column (3) above, we find that the value 6 is repeated the maximum number of times and hence the value of mode is 6 and not 10 which is an irregular item.

In case of continuous frequency distribution, mode is given by the formula :

$$Mode = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2} \qquad ...(2\cdot 7)$$

where l is the lower limit, h the magnitude and f_1 the frequency of the modal class, f_0 and f_2 are the frequencies of the classes preceding and succeeding the modal class respectively.

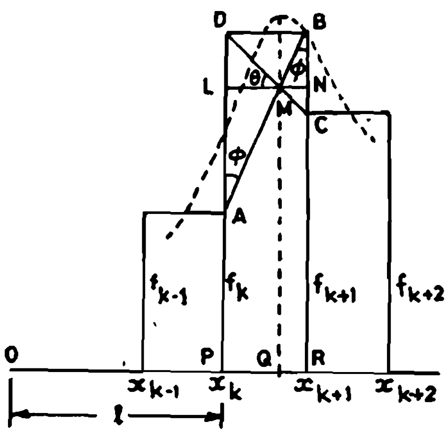
Derivation of the Mode Formula (2·7). Let us consider the continuous frequency distribution :

Class : $x_1 — x_2$, $x_2 — x_3$, ..., $x_k — x_{k+1}$, ..., $x_n — x_{n+1}$

Frequency : f_1 f_2 ... ... f_k ... ... f_n.

If f_k is the maximum of all the frequencies, then the modal class is $(x_k — x_{k+1})$.

Let us further consider a portion of the histogram, namely, the rectangles erected on the modal class and the two adjacent classes. The mode is the value of x for which the frequency curve has a maxima. Let the modal point be Q.



From the figure, we have

$$\tan\theta = \frac{LD}{LM} = \frac{NC}{MN}$$

and
$$\tan\phi = \frac{LM}{AL} = \frac{MN}{NB}$$

∴
$$\frac{LM}{MN} = \frac{LD}{NC} = \frac{AL}{NB} = \frac{AL+LD}{NB+NC} = \frac{AD}{BC}$$

i.e.,
$$\frac{LM}{LN-LM} = \frac{PD-AP}{BR-CR}$$

or
$$\frac{LM}{h-LM} = \frac{f_k - f_{k-1}}{f_k - f_{k+1}},$$ where 'h' is the magnitude of the modal class. Thus solving for LM, we get

$$LM = \frac{h(f_k - f_{k-1})}{(f_k - f_{k+1}) + (f_k - f_{k-1})} = \frac{h(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}$$

Hence Mode $= OQ = OP + PQ = OP + LM$

$$= l + \frac{h(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}$$

Example 2·10. *Find the mode for the following distribution :*

Class - interval :	*0–10*	*10–20*	*20–30*	*30–40*	*40–50*	*50–60*	*60–70*	*70–80*
Frequency :	*5*	*8*	*7*	*12*	*28*	*20*	*10*	*10*

Solution. Here maximum frequency is 28. Thus the class 40–50 is the modal class. Using (2·7), the value of mode is given by

$$\text{Mode} = 40 + \frac{10(28-12)}{(2\times 28 - 12 - 20)} = 40 + 6{\cdot}666 = 46{\cdot}67 \quad \text{(approx.)}$$

Example 2·11. *The Median and Mode of the following wage distribution are known to be Rs. 33·50 and Rs. 34 respectively. Find the values of f_3, f_4 and f_5.*

Wages : (in Rs.)	*0–10*	*10–20*	*20–30*	*30–40*	*40–50*
Frequency :	*4*	*16*	f_3	f_4	f_5

Wages :	*50–60*	*60–70*	*Total*
Frequency :	*6*	*4*	*230*

[Gujarat Univ. B.Sc., 1991]

Solution.

CALCULATIONS FOR MODE AND MEDIAN

Wages (in Rs.)	*Frequency (f)*	*Less than c.f.*
0—10	4	4
10—20	16	20
20—30	f_3	$20 + f_3$
30—40	f_4	$20 + f_3 + f_4$
40—50	f_5	$20 + f_3 + f_4 + f_5$
50—60	6	$26 + f_3 + f_4 + f_5$
60—70	4	$30 + f_3 + f_4 + f_5$
Total	$230 = 30 + f_3 + f_4 + f_5$	

From the above table, we get

$$\Sigma f = 30 + f_3 + f_4 + f_5 = 230$$

$$\Rightarrow \quad f_3 + f_4 + f_5 = 230 - 30 = 200 \quad \ldots(i)$$

Since median is 33·5, which lies in the class 30–40, 30–40 is the median class. Using the median formula, we get

$$Md = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

$$\Rightarrow \qquad 33{\cdot}5 = 30 + \frac{10}{f_4}\left[115 - (20 + f_3)\right]$$

$$\Rightarrow \qquad \frac{33{\cdot}5 - 30}{10} = \frac{95 - f_3}{f_4}$$

$$\Rightarrow \qquad 0{\cdot}35 f_4 = 95 - f_3 \quad \Rightarrow \quad f_3 = 95 - 0{\cdot}35 f_4 \qquad \text{...(ii)}$$

Mode being 34, the modal class is also 30–40. Using mode formula we get :

$$34 = 30 + \frac{10(f_4 - f_3)}{2f_4 - f_3 - f_5}$$

$$\Rightarrow \qquad \frac{34 - 30}{10} = \frac{f_4 + 0{\cdot}35 f_4 - 95}{2f_4 - (200 - f_4)} \qquad \text{[Using } (i) \text{ and } (ii)\text{]}$$

$$\Rightarrow \qquad 0{\cdot}4 = \frac{1{\cdot}35 f_4 - 95}{3f_4 - 200}$$

$$\Rightarrow \qquad 1{\cdot}2 f_4 - 80 = 1{\cdot}35 f_4 - 95$$

$$\Rightarrow \qquad f_4 = \frac{95 - 80}{1{\cdot}35 - 1{\cdot}20} = \frac{15}{0{\cdot}15} = 100 \qquad \text{...(iii)}$$

Substituting in *(ii)* we get :

$$f_3 = 95 - 0{\cdot}35 \times 100 = 60$$

Substituting the values of f_3 and f_4 in *(i)* we get :

$$f_5 = 200 - f_3 - f_4 = 200 - 60 - 100 = 40$$

Hence $\qquad f_3 = 60, f_4 = 100$ and $f_5 = 40$.

Remarks. 1. In case of irregularities in the distribution, or the maximum frequency being repeated or the maximum frequency occurring in the very beginning or at the end of the distribution, the modal class is determined by the method of grouping and the mode is obtained by using (2·7).

Sometimes mode is estimated from the mean and the median. For a symmetrical distribution, mean, median and mode coincide. If the distribution is *moderately asymmetrical,* the mean, median and mode obey the following empirical relationship (due to Karl Pearson) :

$$\text{Mean} - \text{Median} = \tfrac{1}{3}(\text{Mean} - \text{Mode})$$

$$\Rightarrow \qquad \text{Mode} = 3\ \text{Median} - 2\ \text{Mean} \qquad \text{...(2·8)}$$

2. If the method of grouping gives the modal class which does not correspond to the maximum frequency, *i.e.,* the frequency of modal class is not the maximum frequency, then in some situations we may get, $2f_k - f_{k-1} - f_{k+1} = 0$. In such cases, the value of mode can be obtained by the formula :

$$\text{Mode} = l + \frac{h(f_k - f_{k-1})}{|f_k - f_{k-1}| + |f_k - f_{k+1}|}$$

2·7·1. Merits and Demerits of Mode

Merits. *(i)* Mode is readily comprehensible and easy to calculate. Like median, mode can be located in some cases merely by inspection.

(ii) Mode is not at all affected by extreme values.

(iii) Mode can be conveniently located even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude. Open-end classes also do not pose any problem in the location of mode.

Demerits. *(i)* Mode is ill-defined. It is not always possible to find a clearly defined mode. In some cases, we may come across distributions with two modes. Such distributions are called *bi-modal*. If a distribution has more than two modes, it is said to be *multimodal*.

(ii) It is not based upon all the observations.

(iii) It is not capable of further mathematical treatment.

(iv) As compared with mean, mode is affected to a greater extent by fluctuations of sampling.

Uses. Mode is the average to be used to find the ideal size, *e.g.*, in business forecasting, in the manufacture of ready-made garments, shoes, etc.

2·8. Geometric Mean. Geometric mean of a set of n observations is the nth root of their product. Thus the geometric mean G, of n observations $x_i, i = 1, 2, ..., n$ is

$$G = (x_1 . x_2 . \dots . x_n)^{1/n} \qquad ...(2{\cdot}9)$$

The computation is facilitated by the use of logarithms. Taking logarithm of both sides, we get

$$\log G = \frac{1}{n}(\log x_1 + \log x_2 + \dots + \log x_n) = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

$$\therefore \qquad G = \text{Antilog} \left[\frac{1}{n} \sum_{i=1}^{n} \log x_i \right] \qquad ...(2{\cdot}9a)$$

In case of frequency distribution $x_i | f_i$, ($i = 1, 2, ..., n$) geometric mean, G is given by

$$G = \left[x_1^{f_1} . x_2^{f_2} \dots . x_n^{f_n} \right]^{\frac{1}{N}}, \text{ where } N = \sum_{i=1}^{n} f_i \qquad ...(2{\cdot}10)$$

Taking logarithms of both sides, we get

$$\log G = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n)$$

$$= \frac{1}{N} \sum_{i=1}^{n} f_i \log x_i \qquad ...(2{\cdot}10a)$$

Thus we see that logarithm of G is the arithmetic mean of the logarithms of the given values. From (2·10a), we get

$$G = \text{Antilog}\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right) \qquad \text{...(2·10}b\text{)}$$

In the case of grouped or continuous frequency distribution, x is taken to be the value corresponding to the mid-point of the class-intervals.

2·8·1. Merits and Demerits of Geometric Mean

Merits. *(i)* It is rigidly defined.

(ii) It is based upon all the observations.

(iii) It is suitable for further mathematical treatment. If n_1 and n_2 are the sizes, G_1 and G_2 the geometric means of two series respectively, the geometric mean G, of the combined series is given by

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2} \qquad \text{...(2·11)}$$

Proof. Let x_{1i} $(i = 1, 2, ..., n_1)$ and x_{2j} $(j = 1, 2, ..., n_2)$ be n_1 and n_2 items of two series respectively. Then by def.,

$$G_1 = (x_{11} . x_{12} \ldots x_{1n1})^{1/n_1} \quad \Rightarrow \quad \log G_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \log x_{1i}$$

$$G_2 = (x_{21} . x_{22} \ldots x_{2n2})^{1/n_2} \quad \Rightarrow \quad \log G_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \log x_{2j}$$

The geometric mean G of the combined series is given by

$$G = (x_{11} . x_{12} \ldots x_{1n1} . x_{21} . x_{22} \ldots x_{2n2})^{1/(n_1+n_2)}$$

$$\therefore \qquad \log G = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} \log x_{1i} + \sum_{j=1}^{n_2} \log x_{2j}\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 \log G_1 + n_2 \log G_2\right]$$

The result can be easily generalised to more than two series.

(iv) It is not affected much by fluctuations of sampling.

(v) It gives comparatively more weight to small items.

Demerits. *(i)* Because of its abstract mathematical character, geometric mean is not easy to understand and to calculate for a non-mathematics person.

(ii) If any one of the observations is zero, geometric mean becomes zero and if any one of the observations is negative, geometric mean becomes imaginary regardless of the magnitude of the other items.

Uses. Geometric mean is used –

(i) To find the rate of population growth and the rate of interest.

(ii) In the construction of index numbers.

Example 2·12. *Show that in finding the arithmetic mean of a set of readings on thermometer it does not matter whether we measure temperature in Centigrade or Fahrenheit, but that in finding the geometric mean it does matter which scale we use.* **[Patna Univ. B.Sc., 1991]**

Solution. Let $C_1, C_2, ..., C_n$ be the n readings on the Centigrade thermometer. Then their arithmetic mean $\overline{C}$ is given by :

$$\overline{C} = \frac{1}{n}(C_1 + C_2 + \ldots + C_n)$$

If F and C be the readings in Fahrenheit and Centigrade respectively then we have the relation :

$$\frac{F-32}{180} = \frac{C}{100} \quad \Rightarrow \quad F = 32 + \frac{9}{5}C.$$

Thus the Fahrenheit equivalents of $C_1, C_2, ..., C_n$ are

$$32 + \frac{9}{5}C_1,\ 32 + \frac{9}{5}C_2,\ \ldots,\ 32 + \frac{9}{5}C_n,$$

respectively.

Hence the arithmetic mean of the readings in Fahrenheit is

$$\begin{aligned}\overline{F} &= \frac{1}{n}\left\{\left(32 + \frac{9}{5}C_1\right) + \left(32 + \frac{9}{5}C_2\right) + \ldots + \left(32 + \frac{9}{5}C_n\right)\right\}\\ &= \frac{1}{n}\left\{32n + \frac{9}{5}(C_1 + C_2 + \ldots + C_n)\right\}\\ &= 32 + \frac{9}{5}\left(\frac{C_1 + C_2 + \ldots + C_n}{n}\right)\\ &= 32 + \frac{9}{5}\overline{C}.\end{aligned}$$

which is the Fahrenheit equivalent of $\overline{C}$.

Hence in finding the arithmetic mean of a set of n readings on a thermometer, it is immaterial whether we measure temperature in Centigrade or Fahrenheit.

Geometric mean G, of n readings in Centigrade is

$$G = (C_1 . C_2 \ldots C_n)^{1/n}$$

Geometric mean G_1, (say), of Fahrenheit equivalents of $C_1, C_2, ..., C_n$ is

$$G_1 = \left\{\left(32 + \frac{9}{5}C_1\right)\left(32 + \frac{9}{5}C_2\right)\ldots\left(32 + \frac{9}{5}C_n\right)\right\}^{1/n}$$

which is not equal to Fahrenheit equivalent of G, *viz.*,

$$\left\{\frac{9}{5}(C_1 . C_2 \ \ldots \ C_n)^{1/n} + 32\right\}$$

Hence in finding the geometric mean of the n readings on a thermometer, the scale (Centrigrade or Fahrenheit) is important.

2·9. Harmonic Mean. Harmonic mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocals of the given values. Thus, harmonic mean H, of n observations x_i, $i = 1, 2, ..., n$ is

$$H = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} (1/x_i)} \qquad ...(2·12)$$

In case of frequency distribution $x_i \mid f_i$, $(i = 1, 2, ..., n)$,

$$H = \frac{1}{\frac{1}{N} \sum_{i=1}^{n} (f_i/x_i)}, \quad \left[N = \sum_{i=1}^{n} f_i\right] \qquad ...(2·12a)$$

2·9·1. Merits and Demerits of Harmonic Mean

Merits. Harmonic mean is rigidly defined, based upon all the observations and is suitable for further mathematical treatment. Like geometric mean, it is not affected much by fluctuations of sampling. It gives greater importance to small items and is useful only when small items have to be given a greater weightage.

Demerits. Harmonic mean is not easily understood and is difficult to compute.

Example 2·13. *A cyclist pedals from his house to his college at a speed of 10 m.p.h. and back from the college to his house at 15 m.p.h. Find the average speed.*

Solution. Let the distance from the house to the college be x miles. In going from house to college, the distance (x miles) is covered in $\frac{x}{10}$ hours, while in coming from college to house, the distance is covered in $\frac{x}{15}$ hours. Thus a total distance of $2x$ miles is covered in $\left(\frac{x}{10} + \frac{x}{15}\right)$ hours.

$$\text{Hence average speed} = \frac{\text{Total distance travelled}}{\text{Total time taken}} = \frac{2x}{\left(\frac{x}{10} + \frac{x}{15}\right)}$$

$$= \frac{2}{\left(\frac{1}{10} + \frac{1}{15}\right)} = 12 \text{ } m.p.h.$$

Remark. 1. In this case the average speed is given by the harmonic mean of 10 and 15 and not by the airthmetic mean.

Rather, we have the following general result :

If equal distances are covered (travelled) per unit of time with speeds equal to $V_1, V_2, ..., V_n$, say, then the average speed is given by the harmonic mean of $V_1, V_2, ..., V_n$, *i.e.*,

$$\text{Average speed} = \frac{n}{\left(\frac{1}{V_1} + \frac{1}{V_2} + ... + \frac{1}{V_n}\right)} = \frac{n}{\Sigma \left(\frac{1}{V}\right)}$$

Proof is left as an exercise to the reader.

Hint. $\text{Speed} = \dfrac{\text{Distance}}{\text{Time}} \Rightarrow \text{Time} = \dfrac{\text{Distance}}{\text{Speed}}$

$$\text{Average Speed} = \frac{\text{Total distance travelled}}{\text{Total time taken}}$$

2. Weighted Harmonic Mean. Instead of fixed (constant) distance being travelled with varying speed, let us now suppose that different distances, say, $S_1, S_2, ..., S_n$, are travelled with different speeds, say, $V_1, V_2, ..., V_n$ respectively. In that case, the average speed is given by the weighted harmonic mean of the speeds, the weights being the corresponding distances travelled, *i.e.*,

$$\text{Average speed} = \frac{S_1 + S_2 + \ldots + S_n}{\left(\dfrac{S_1}{V_1} + \dfrac{S_2}{V_2} + \ldots + \dfrac{S_n}{V_n}\right)} = \frac{\Sigma S}{\Sigma \left(\dfrac{S}{V}\right)}$$

Example 2·14. *You can take a trip which entails travelling 900 km. by train at an average speed of 60 km. per hour, 3000 km. by boat at an average of 25 km. p.h., 400 km. by plane at 350 km. per hour and finally 15 km. by taxi at 25 km. per hour. What is your average speed for the entire distance ?*

Solution. Since different distances are covered with varying speeds, the required average speed for the entire distance is given by the weighted harmonic mean of the speeds (in km.p.h.), the weights being the corresponding distances covered (in kms.).

COMPUTATION OF WEIGHTED H. M.

Speed (km. / hr.) X	*Distance (in km.)* W	W/X
60	900	15·00
25	3000	120·00
350	400	1·43
25	15	0·60
Total	$\Sigma W = 4315$	$\Sigma (W/X) = 137·03$

Average speed

$$= \frac{\Sigma W}{\Sigma (W/X)} = \frac{4315}{137·03} = 31·489 \text{ km.p.h.}$$

2·10. Selection of an Average. From the preceding discussion it is evident that no single average is suitable for all practical purposes. Each one of the average has its own merits and demerits and thus its own particular field of importance and utility. We cannot use the averages indiscriminately. A judicious selection of the average depending on the nature of the data and the purpose of the enquiry is essential for sound statistical analysis. Since arithmetic mean satisfies all the properties of an ideal average as laid down by Prof. Yule, is familiar to a layman and further has wide applications in statistical theory at large, it may be regarded as the best of all the averages.

2·11. Partition Values. These are the values which divide the series into a number of equal parts.

The three points which divide the series into four equal parts are called *quartiles.* The first, second and third points are known as the first, second and third quartiles respectively. The first quartile, Q_1, is the value which exceed 25% of the observations and is exceeded by 75% of the observations. The second quartile, Q_2, coincides with median. The third quartile, Q_3, is the point which has 75% observations before it and 25% observations after it.

The nine points which divide the series into ten equal parts are called *deciles* whereas *percentiles* are the ninety-nine points which divide the series into hundred equal parts. For example, D_7, the seventh decile, has 70% observations before it and P_{47}, the forty-seventh percentile, is the point which exceed 47% of the observations. The methods of computing the partition values are the same as those of locating the median in the case of both discrete and continuous distributions.

Example 2·15. *Eight coins were tossed together and the number of heads resulting was noted. The operation was repeated 256 times and the frequencies (f) that were obtained for different values of x, the number of heads, are shown in the following table. Calculate median, quartiles, 4th decile and 27th precentile.*

x :	0	1	2	3	4	5	6	7	8
f :	1	9	26	59	72	52	29	7	1

Solution.

x :	0	1	2	3	4	5	6	7	8
f :	1	9	26	59	72	52	29	7	1
c.f. :	1	10	36	95	167	219	248	255	256

Median : Here $N/2 = 256/2 = 128$. Cumulative frequency (*c.f.*) just greater than 128 is 167. Thus, median = 4.

Q_1 : Here $N/4 = 64$. *c.f.* just greater than 64 is 95. Hence, $Q_1 = 3$.

Q_3 : Here $3N/4 = 192$ and *c.f.* just greater than 192 is 219. Thus $Q_3 = 5$.

D_4 : $\dfrac{4N}{10} = 4 \times 25{\cdot}6 = 102{\cdot}4$ and *c.f.* just greater than 102·4 is 167. Hence $D_4 = 4$.

P_{27} : $\dfrac{27N}{100} = 27 \times 2{\cdot}56 = 69{\cdot}12$ and *c.f.* just greater than 69·12 is 95. Hence $P_{27} = 3$.

2·11·1. Graphical Location of the Partition Values. The partition values, *viz.*, quartiles, deciles and percentiles, can be conveniently located with the help of a curve called the 'cumulative frequency curve' or 'Ogive'. The procedure is illustrated below.

First form the cumulative frequency table. Take the class intervals (or the variate values) along the *x*–axis and plot the corresponding cumulative frequencies along the *y*–axis against the *upper limit* of the class interval (or against the variate value in the case of discrete frequency distribution). The curve obtained on joining the points so obtained by means of free hand drawing is called the *cumulative frequency curve* or *ogive*. The graphical location of partition values from this curve is explained below by means of an example.

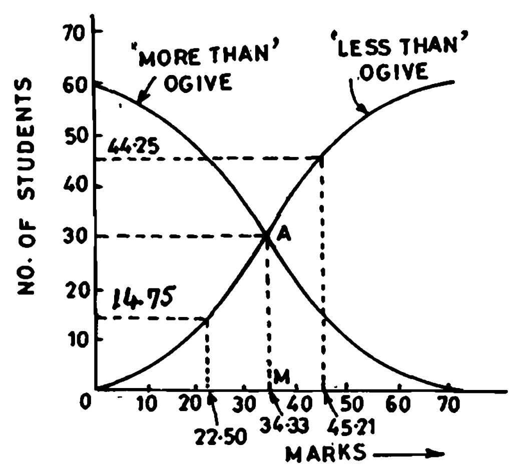
Example 2·16. *Draw the cumulative frequency curve for the following distribution showing the number of marks of 59 students in Statistics.*

Marks-group :	0—10	10—20	20—30	30—40	40—50	50—60	60—70
No. of Students :	4	8	11	15	12	6	3

Solution.

Marks-group	*No. of Students*	*Less than c.f.*	*More than c.f.*
0—10	4	4	59
10—20	8	12	55
20—30	11	23	47
30—40	15	38	36
40—50	12	50	21
50—60	6	56	9
60—70	3	59	3

Taking the marks–group along x-axis and *c.f.* along y-axis, we plot the cumulative frequencies, *viz.*, 4, 12, 23, ..., 59 against the upper limits of the corresponding classes, *viz.*, 10, 20, ..., 70 respectively. The smooth curve obtained on joining these points is called *ogive* or more particularly *'less than' ogive.*



If we plot the 'more than' cumulative frequencies, *viz.*, 59, 55,..., 3 against the lower limits of the corresponding classes, *viz.*, 0, 10, ..., 60 and join the points by a smooth curve, we get cumulative frequency curve which is also known as *ogive* or more particularly *'more than' ogive.*

To locate graphically the value of median, mark a point corresponding to $N/2$ along y-axis. At this point draw a line parallel to x-axis meeting the ogive at the point 'A' (say). From 'A' draw a line prependicular to x-axis meeting it in 'M' (say). Then abscissa of 'M' gives the value of median.

To locate the values of Q_1 (or Q_3), we mark the points along y-axis corresponding to $N/4$ (or $3N/4$) and proceed exactly similarly.

In the above example, we get from ogive

Median= 34·33, Q_1 = 22·50, and Q_3 = 45·21.

Remarks. 1. The median can also be located as follows :

From the point of intersection of 'less than' ogive and 'more than' ogive, draw perpendicular to OX. The abscissa of the point so obtained gives median.

2. Other partition values, *viz.*, deciles and percentiles, can be similarly located from 'ogive'.

EXERCISE

1. *(a)* What are grouped and ungrouped frequency distributions? What are their uses? What are the considerations that one has to bear in mind while forming the frequency distribution?

(b) Explain the method of constructing Histogram and Frequency Polygon. Which, out of these two, is better representative of frequencies of *(i)* a particular group, and *(ii)* whole group.

2. What are the principles governing the choice of :

(i) Number of class intervals,

(ii) The length of the class interval,

(iii) The mid-point of the class interval.

3. Write short notes on :

(i) Frequency distribution,

(ii) Histogram, frequency polygon and frequency curve,

(iii) Ogive.

4. *(a)* What are the properties of a good average? Examine these properties with reference to the Arithmetic Mean, the Geometric Mean and the Harmonic Mean, and give an example of situations in which each of them can be the appropriate measure for the average.

(b) Compare mean, median and mode as measures of location of a distribution.

(c) The mean is the most common measure of central tendency of the data. It satisfies almost all the requirements of a good average. The median is also an average, but it does not statisfy all the requirements of a good average. However, it carries certain merits and hence is useful in particular fields. Critically examine both the averages.

(d) Describe the different measures of central tendency of a frequency distribution, mentioning their merits and demerits.

5. Define *(i)* arithmetic mean, *(ii)* geometric mean and *(iii)* harmonic mean of grouped and ungrouped data. Compare and contrast the merits and demerits of them. Show that the geometric mean is capable of further mathematical treatment.

6. *(a)* When is an average a meaningful statistics? What are the requisites of a statisfactory average? In this light compare the relative merits and demerits of three well-known averages.

(b) What are the chief measures of central tendency? Discuss their merits.

7. Show that *(i)* Sum of deviations about arithmetic mean is zero.

(ii) Sum of absolute deviations about median is least.

(iii) Sum of the squares of deviations about arithmetic mean is least.

8. The following numbers give the weights of 55 students of a class. Prepare a suitable frequency table.

42	74	40	60	82	115	41	61	75	83	63
53	110	76	84	50	67	65	78	77	56	95
68	69	104	80	79	79	54	73	59	81	100
66	49	77	90	84	76	42	64	69	70	80
72	50	79	52	103	96	51	86	78	94	71

(i) Draw the histogram and frequency polygon of the above data.

(ii) For the above weights, prepare a cumulative frequency table and draw the less than ogive.

9. *(a)* What are the points to be borne in mind in the formation of frequency table?

Choosing appropriate class-intervals, form a frequency table for the following data:

10·2	0·5	5·2	6·1	3·1	6·7	8·9	7·2	8·9
5·4	3·6	9·2	6·1	7·3	2·0	1·3	6·4	8·0
4·3	4·7	12·4	8·6	13·1	3·2	9·5	7·6	4·0
5·1	8·1	1·1	11·5	3·1	6·8	7·0	8·2	2·0
3·1	6·5	11·2	12·0	5·1	10·9	11·2	8·5	2·3
3·4	5·2	10·7	4·9	6·2				

(b) What are the considerations one has to bear in mind while forming a frequency distribution?

A sample consists of 34 observations recorded correct to the nearest integer, ranging in value from 201 to 337. If it is decided to use seven classes of width 20 integers and to begin the first class at 199·5, find the class limits and class marks of the seven classes.

(c) The class marks in a frequency table (of whole numbers) are given to be 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50. Find out the following :

(i) the true classes.

(ii) the true class limits.

(iii) the true upper class limits.

10. *(a)* The following table shows the distribution of the number of students per teacher in 750 colleges :–

Students :	1	4	7	10	13	16	19	22	25	28
Frequency :	7	46	165	195	189	89	28	19	9	3

Draw the histogram for the data and superimpose on it the frequency polygon.

(b) Draw the histogram and frequency curve for the following data.

Monthly wages in Rs.	10–13	13–15	15–17	17–19	19–21	21–23	23–25
No. of workers	6	53	85	56	21	16	8

(c) Draw a histogram for the following data :

Age (in years) :	2–5	5–11	11–12	12–14	14–15	15–16
No. of boys :	6	6	2	5	1	3

11. *(a)* Three people A, B, C were given the job of finding the average of 5000 numbers. Each one did his own simplification. A's method : Divide the sets into sets of 1000 each, calculate the average in each set and then calculate the average of these averages. B's method : Divide the set into 2,000 and 3,000 numbers, take average in each set and then take the average of the averages. C's method :500 numbers were unities. He averaged all other numbers and then added one. Are these methods correct?

Ans. Correct, not correct, not correct.

(b) The total sale (in '000 rupees) of a particular item in a shop, on 10 consecutive days, is reported by a clerk as, 35·00, 29·60, 38·00, 30·00, 40·00, 41·00, 42·00, 45·00, 3·60, 3·80. Calculate the average. Later it was found that there was a number 10·00 in the machine and the reports of 4th to 8th days were 10·00 more than the true values and in the last 2 days he put a decimal in the wrong place thus for example 3·60 was really 36·0. Calculate the true mean value.

Ans. 30·8, 32·46.

12. *(a)* Given below is the distribution of 140 candidates obtaining marks X or higher in a certain examination (all marks are given in whole numbers) :

X :	10	20	30	40	50	60	70	80	90	100
$c.f.$:	140	133	118	100	75	45	25	9	2	0

Calculate the mean, median and mode of the distribution.

Hint.

Class	Frequency (f)	Class boundaries	Mid value	$c.f.$ (less than)
10—19	140 – 133 = 7	9·5—19·5	14·5	7
20—29	133 – 118 = 15	19·5—29·5	24·5	22
30—39	118 – 100 = 18	29·5—39·5	34·5	40
40—49	100 – 75 = 25	39·5—49·5	44·5	65
50—59	75 – 45 = 30	49·5—59·5	54·5	95
60—69	45 – 25 = 20	59·5—69·5	64·5	115
70—79	25 – 9 = 16	69·5—79·5	74·5	131
80—89	9 – 2 = 7	79·5—89·5	84·5	138
90—99	2 – 0 = 2	89·5—99·5	94·5	140

$$\text{Mean} = 54.5 + \frac{10 \times (-53)}{140} = 50.714$$

$$\text{Median} = 49.5 + \frac{10}{30}\left(\frac{140}{2} - 65\right) = 51.167$$

(b) The four parts of a distribution are as follows :

Part	*Frequency*	*Mean*
1	50	61
2	100	70
3	120	80
4	30	83

Find the mean of the distribution. **(Madurai Univ. B.Sc., 1988)**

13. *(a)* Define a 'weighted mean'. If several sets of observations are combined into a single set, show that the mean of the combined set is the weighted mean of several sets.

(b) The weighted geometric mean of three numbers 229, 275 and 125 is 203 The weights for the first and second numbers are 2 and 4 respectively. Find the weight of third. Ans. 3.

14. Define the weighted arithmetic mean of a set of numbers. Show that it is unaffected if all weights are multiplied by some common factor.

The following table shows some data collected for the regions of a country:

Region	*Number of inhabitants (million)*	*Percentage of literates*	*Average annual income per person (Rs.)*
A	10	52	850
B	5	68	620
C	18	39	730

Obtain the overall figures for the three regions taken together. Prove the formulae you use. **[Calcutta Univ. B.A.(Hons.), 1991]**

15. Draw the Ogives and hence estimate the median.

Class	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79
Frequency	8	32	142	216	240	206	143	13

16. The following data relate to the ages of a group of workers in a factory.

Ages	*No. of workers*	*Ages*	*No. of workers*
20—25	35	40—45	90
25—30	45	45—50	74
30—35	70	50—55	51
35—40	105	55—60	30

Draw the percentage cumulative curve and find from the graph the number of workers between the ages 28—48.

17. *(a)* The mean of marks obtained in an examination by a group of 100 students was found to be 49·96. The mean of the marks obtained in the same examination by another group of 200 students was 52·32. Find the mean of the marks obtained by both the groups of students taken together.

(b) A distribution consists of three components with frequencies 300, 200 and 600 having their means 16, 8 and 4 respectively. Find the mean of the combined distribution.

(c) The mean marks got by 300 students in the subject of Statistics are 45. The mean of the top 100 of them was found to be 70 and the mean of the last 100 was known to be 20. What is the mean of the remaining 100 students?

(d) The mean weight of 150 students in a certain class is 60 kilograms. The mean weight of boys in the class is 70 kilograms and that of the girls is 55 kilograms. Find the number of boys and number of girls in the class.

Ans. *(a)* 51·53, *(b)* 8, *(c)* 45, *(d)* Boys = 50, Girls = 100.

18. From the following data, calculate the percentage of workers getting wages

(a) more than Rs. 44, *(b)* between Rs. 22 and Rs. 58, *(c)* Find Q_1 and Q_3.

Wages (Rs.)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
No. of workers	20	45	85	160	70	55	35	30

Hint. Assuming that frequencies are uniformly distributed over the entire interval,

(a) Number of persons with wages more than Rs. 44 is

$$\left(\frac{50-44}{10}\times 70\right)+55+35+30=162$$

Hence the percentage of workers getting over Rs. 44 is

$$=\frac{162}{500}\times 100=32{\cdot}4\%$$

(b) Percentage of workers getting wages between Rs. 22 and Rs. 58 is

$$\left[\left(\frac{30-22}{10}\times 85\right)+160+70+\left(\frac{58-50}{10}\times 55\right)\right]\times 100+500=68{\cdot}4\%$$

19. For the two frequency distributions give below the mean calculated from the first was 25·4 and that from the second was 32·5. Find the values of x and y.

Class	*Distribution I Frequency*	*Distribution II Frequency*
10—20	20	4
20—30	15	8
30—40	10	4
40—50	x	$2x$
50—60	y	y

Ans. $x=3$, $y=2$

20. A number of particular articles has been classified according to their weights. After drying for two weeks the same articles have again been weighted and similarly classified. It is known that the median weight in the first weighing was 20·83 oz. while in the second weighing it was 17·35 oz. Some frequencies a and b in the first weighing and x and y in the second are missing. It is known that $a = \frac{1}{3}x$ and $b = \frac{1}{2}y$. Find out the values of the missing frequencies.

Class	*Frequencies*		*Class*	*Frequencies*	
	1st weighing	*IInd weighing*		*1st weighing*	*IInd weighing*
0—5	a	x	15—20	52	50
5—10	b	y	20—25	75	30
10—15	11	40	25—30	22	28

Hint. We have $x = 3a,\ y = 2b$,

$N_1 =$ Total frequency in 1st weighing $= 160 + a + b$.

$N_2 =$ Total frequency in 2nd weighing $= 148 + x + y = 148 + 3a + 2b$.

Using Median formula, we shall get

$$20{\cdot}83 = 20 + \frac{5}{75}\left[\frac{N_1}{2} - (63 + a + b)\right]$$

$$\Rightarrow \quad 15\,(20{\cdot}83 - 20) = \frac{160 + a + b}{2} - (63 + a + b)$$

$$\Rightarrow \quad 12{\cdot}45 = 17 - \frac{a+b}{2}$$

$$\Rightarrow \quad a + b = 2\,(17 - 12{\cdot}45) = 9{\cdot}10 \simeq 9 \qquad \ldots(*)$$

Since a and b, being frequencies are integral valued, $a + b$ is also integral valued. Now the median of 2nd weighing gives :

$$17{\cdot}35 = 15 + \frac{5}{50}\left[\frac{148 + 3a + 2b}{2} - (40 + x + y)\right]$$

$$\Rightarrow \quad 10 \times 2{\cdot}35 = 74 + \frac{3a + 2b}{2} - 40 - 3a - 2b$$

$$\Rightarrow \quad \frac{3a + 2b}{2} = 34 - 23{\cdot}5 = 10{\cdot}5$$

$$\Rightarrow \quad 3a + 2b = 21 \qquad \ldots(**)$$

Multiplying $(*)$ by 3, we get

$$3a + 3b = 27 \qquad \ldots(***)$$

Subtracting $(**)$ from $(***)$, we get $b = 6$. Substituting in $(*)$, we get $a = 9 - 6 = 3$.

$$\therefore \quad a = 3,\ b = 6\ ;\ \ x = 3a = 9,\ y = 2b = 12.$$

21. From the following table showing the wage distribution in a certain factory, determine:

(*a*) the mean wage,

(*b*) the median wage,

(*c*) the modal wage,

(*d*) the wage limits for the middle 50% of the wage earners,

(*e*) the percentages of workers who earned between Rs.75 and Rs.125.

(*f*) the percentage who earned more than Rs.150 per week, and

(*g*) the percentage who earned less than Rs.100 per week.

Weekly wages (Rs.)	*No. of employees*	*Weekly wages (Rs.)*	*No. of employees*
20–40	8	120–140	35
40–60	12	140–160	18
60–80	20	160–180	7
80–100	30	180–200	5
100–120	40		

Ans. (*a*) $\overline{X} = 108{\cdot}5$, (*b*) Med. = 108·75, (*c*) Mo = 118·3, (*d*) 81·25, 129·3 (*e*) 48, (*f*) 12, (*g*) 40.

22. (*a*) Explain how the ogives are drawn for any frequency distribution. Point out the method of finding out the values of median, mode, quartiles, deciles and percentiles graphically. Also, write down the formula for the computation of each of them for any frequency distribution.

(*b*) The following table gives the frequency distribution of marks in a class of 65 students.

Marks	*No. of Students*	*Marks*	*No. of students*
0–4	10	14–18	5
4–8	12	18–20	3
8–12	18	20–25	4
12–14	7	25 and over	6
Total			65

Calculate : (*i*) Upper and lower quartiles.

(*ii*) No. of students who secured marks more than 17.

(*iii*) No. of students who secured marks between 10 and 15.

(*c*) The following table shows the age distribution of heads of families in a certain country during the year 1957. Find the median, the third quartile and the second decile of the distribution. Check your results by the graphical method.

Age of head of family years	Under 25	25–29	30–34	35–44	45–54	55–64	65–74	above 74
Number (million)	2·3	4·1	5·3	10·6	9·7	6·8	4·4	1·8
								Total 45

Ans. Md = 45·2 yrs.; $Q_3 = 57{\cdot}5$ yrs.; $D_2 = 32{\cdot}5$ yrs.

23. The following data represent travel expenses (other than transportation) for 7 trips made during November by a salesman for a small firm :

Trip	*Days*	*Expense (Rs.)*	*Expense per day (Rs.)*
1	0·5	13·50	27
2	2·0	12·00	6
3	3·5	17·50	5
4	1·0	9·00	9
5	9·0	27·00	3
6	0·5	9·00	18
7	8·5	17·00	2
Total	25·0	105·00	70

An auditor criticised these expenses as excessive, asserting that the average expense per day is Rs. 10 (Rs. 70 divided by 7). The salesman replied that the average is only Rs. 4·20 (Rs. 105 divided by 25) and that in any event the median is the appropriate measure and is only Rs. 3. The auditor rejoined that the arithmetic mean is the appropriate measure, but that the median is Rs. 6.

You are required to :

(a) Explain the proper interpretation of each of the four averages mentioned.

(b) Which average seems appropriate to you ?

24. *(a)* Define Geometric and Harmonic means and explain their uses in statistical analysis.

You take a trip which entails travelling 900 miles by train at an average speed of 60 m.p.h., 300 miles by boat at an average of 25 m.p.h., 400 miles by plane at 350 m.p.h. and finally 15 miles by taxi at 25 m.p.h. What is your speed for the entire distance?

(b) A train runs 25 miles at a speed of 30 m.p.h., another 50 miles at a speed of 40 m.p.h., then due to repairs of the track travels for 6 minutes at a speed of 10 m.p.h. and finally covers the remaining distance of 24 miles at a speed of 24 m.p.h. What is the average speed in m.p.h.?

(c) A man motors from A to B. A large part of the distance is uphill and he gets a mileage of only 10 per gallon of gasoline. On the return trip he makes 15 miles per gallon. Find the harmonic mean of his mileage. Verify the fact that this is the proper average to be used by assuming that the distance from A to B is 60 miles.

(d) Calculate the average speed of a car running at the rate of 15 km.p.h. during the first 30 kms., at 20 km.p.h. during the second 30 kms. and at 25 kmp.h. during the third 30 kms.

25. The following table shows the distribution of 100 families according to their expenditure per week. Number of families corresponding to expenditure groups Rs. (10—20) and Rs.(30—40) are missing from the table. The median and

mode are given to be Rs.25 and 24 Calculate the missing frequencies and then arithmetic mean of the data :

Expenditure :	0—10	10—20	20—30	30—40	40—50
No. of families :	14	?	27	?	15

Hint.

Expenditure	*No. of Families*	*Cumulative frequencies*
0—10	14	14
10—20	f_1	$14+f_1$
20—30	27	$41+f_1$
30—40	f_2	$41+f_1+f_2$
40—50	15	$56+f_1+f_2$

$$\therefore \qquad 25 = 20 + \frac{\dfrac{56+f_1+f_2}{2} - (14+f_1)}{27} \times 10$$

and
$$24 = 20 + \frac{27-f_1}{2\times 27 - f_1 - f_2} \times 10$$

Simplifying these equations, we get

$$f_1 - f_2 = 1$$

and
$$3f_1 - 2f_2 = 27.$$

Ans. 25, 24

26. *(a)* The numbers 3·2, 5·8, 7·9 and 4·5 have frequencies x, $(x+2)$, $(x-3)$ and $(x+6)$ respectively. If their arithmetic mean is 4·876, find the value of x.

(b) If $M_{g.x}$ is the geometric mean of N x's amd $M_{g.y}$ is the geometric mean of N y's, then the geometric mean M_g of the $2N$ values is given by

$$M_g^2 = M_{g.x} M_{g.y}. \qquad \textbf{(Nagpur Univ. B.Sc., 1990)}$$

(c) The weighted geometric mean of the three numbers 229, 275 and 125 is 203. The weights for the first and the second numbers are 2 and 4 respectively. Find the weight of the third. **Ans.** 3.

27. The geometric mean of 10 observations on a certain variable was calculated as 16·2. It was later discovered that one of the observations was wrongly recorded as 12·9; in fact it was 21·9. Apply appropriate correction and calculate the correct geometric mean.

Hint. Correct value of the geometic mean, G' is given by

$$G' = \left(\frac{(16\cdot2)^{10} \times 21\cdot9}{12\cdot9} \right)^{1/10} = 17\cdot68$$

28. A variate takes the values $a, ar, ar^2, \ldots, ar^{n-1}$ each with frequency unity. If A, G and H are respectively the arithmetic mean, geometric mean and harmonic mean, show that

$$A = \frac{a(1-r^n)}{n(1-r)},\ G = ar^{(n-1)/2},\ H = \frac{an(1-r)r^{n-1}}{(1-r^n)}$$

Prove that $G^2 = AH$. Prove also that $A > G > H$ unless $r = 1$, when all the three means coincide.

29. If $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{x}_2 = \frac{1}{n}\sum_{i=2}^{n+1} x_i$ and $\bar{x}_3 = \frac{1}{n}\sum_{i=3}^{n+2} x_i$

then show that

(*a*) $\bar{x}_2 = \bar{x}_1 + \frac{1}{n}(x_{n+1} - x_1)$, and (*b*) $\bar{x}_3 = \bar{x}_2 + \frac{1}{n}(x_{n+2} - x_2)$

30. A distribution $x_1, x_2, \ldots, x_n$ with frequencies $f_1, f_2, \ldots, f_n$ transformed into the distribution $X_1, X_2, \ldots, X_n$ with the same corresponding frequencies by the relation $X_r = ax_r + b$, where a and b are constants. Show that the mean, median and mode of the new distribution are given in terms of those of the first distribution by the same transformation. **[Kanpur Univ. B.Sc., 1992]**

Use the method indicated above to find the mean of the following distribution: x (duration of telephone conversation in seconds)

49·5,	149·5,	249·5,	349·5,	449·5,	549·5,	649·5,	749·5,	849·5,	949·5

f (respective frequency)

6	28	88	180	247	260	132	48	11	5

31. If $\bar{x}_w$ is the weighted mean of x_i's with weights w_i, prove that

$$\left(\sum_{i=1}^{n} w_i\right)\left(\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2\right) = \sum_{i=1}^{n}\sum_{j>i}^{n} w_i w_j (x_i - x_j)^2, \text{ where } \sum_{i=1}^{n} w_i \neq 0.$$

(Allahabad Univ. B.Sc., 1992)

Hint. $\left[\sum_{i=1}^{n}\sum_{j>i} w_i w_j (x_i - x_j)^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=i}^{n} w_i w_j (x_i - x_j)^2\right.$

$$\left. = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=i}^{n} w_i w_j \left\{(x_i - \bar{x}_w) - (x_j - \bar{x}_w)\right\}^2\right]$$

32. In a frequency table, the upper boundary of each class interval has a constant ratio to the lower boundary. Show that the geometric mean G may be expressed by the formula :

$$\log G = x_0 + \frac{c}{N}\sum_i f_i (i-1)$$

where x_0 is the logarithm of the mid-value of the first interval and c is the logarithm of the ratio between upper and lower boundaries.

[Delhi Univ. B.Sc. (Stat. Hons.), 1990, 1986]

33. Find the minimum value of :

(i) $f(x) = (x-6)^2 + (x+3)^2 + (x-8)^2 + (x+4)^2 + (x-3)^2$

(ii) $g(x) = |x-6| + |x+3| + |x-8| + |x+4| + |x-3|$.

[Delhi Univ. B.Sc.(Stat. Hons.), 1991]

Hint. The sum of squares of deviations is minimum when taken from arithmetic mean and the sum of absolute deviations is minimum when taken from median.

34. If A, G and H be the arithmetic mean, geometric mean and harmonic mean respectively of two positive numbers a and b, then prove that :

(i) $A \geq G \geq H$.

When does the equality sign hold?

(ii) $G^2 = AH$.

35 Calculate simple and weighted arithmetic averages from the following data and comment on them :

Designation	*Monthly salary (in Rs.)*	*Strength of the cadre*
Class I Officers	1,500	10
Class II Officers	800	20
Subordinate staff	500	70
Clerical staff	250	100
Lower staff	100	150

Ans. $\overline{X}$ = Rs. 630, $\overline{X}_w$ = Rs. 302·86. Latter is more representative.

36. Treating the number of letters in each word in the following passage as the variable x, prepare the frequency distribution table and obtain its mean, median, mode.

"The reliability of data must always be examined before any attempt is made to base conclusions upon them. This is true of all data, but particularly so of numerical data, which do not carry their quality written large on them. It is a waste of time to apply the refined theoretical methods of Statistics to data which are suspect from the beginning."

Ans. Mean = 4·565, Median = 4, Mode = 3.

OBJECTIVE TYPE QUESTIONS

I. Match the correct parts to make a valid statement :

(a) Arithmetic Mean	*(i)* $l + [f_2/(f_1 + f_2)] \times i$
(b) Geometric Mean	*(ii)* $(x_1 . x_2 . \ldots . x_n)^{1/n}$
(c) Harmonic Mean	*(iii)* $\Sigma fX / \Sigma f$
(d) Median	*(iv)* $l + \dfrac{N/2 - c.f.}{f} \times i$

(e) Mode (v) $\left[\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}\right)\right]^{-1}$

(vi) $l=\frac{f_1-f_o}{2f_1-f_o-f_2}\times i$

II. Which measure of location will be suitable to compare:

(i) heights of students in two classes;
(ii) size of agricultural holdings;
(iii) average sales for various years;
(iv) intelligence of students;
(v) per capita income in several countries;
(vi) sale of shirts with collar size; 16″, $15\frac{1}{2}$″, 15″, 14″, 13″, 15″;
(vii) marks obtained 10, 8, 12, 4, 7, 11, and X ($X < 5$).

Ans. (i) Mean, (ii) Mode, (iii) Mean, (iv) Median, (v) Mean, (vi) Mode, (vii) Median.

III. Which of the following are true for all sets of data?

(i) Arithmetic Mean ≤ median ≤ mode,
(ii) Arithmetic mean ≥ median ≥ mode,
(iii) Arithmetic mean = median = mode
(iv) None of these.

IV. Which of the following are true in respect of any distribution?

(i) The percentile points are in the ascending order.
(ii) The percentile points are equispaced.
(iii) The median is the mid-point of the range and the distribution.
(iv) A unique median value exists for each and every distribution.

V. Find out the missing figures :

(a) Mean = ? (3 Median – Mode).
(b) Mean – Mode = ? (Mean – Median).
(c) Median = Mode + ? (Mean – Mode).
(d) Mode = Mean – ? (Mean – Median).

Ans. (a) 1/2, (b) 3, (c) 2/3, (d) 3.

VI. Fill in blanks :

(i) Harmonic mean of a number of observations is ......

(ii) The geometric mean of 2, 4, 16, and 32 is ......

(iii) The strength of 7 colleges in a city are 385; 1,748; 1,343; 1,935; 786; 2,874 and 2,108. Then the median strength is ......

(iv) The geometric mean of a set of values lies between arithmetic mean and...

(v) The mean and median of 100 items are 50 and 52 respectively. The value of the largest item is 100. It was later found that it is actually 110. Therefore, the true mean is ... and the true median is ......

(vi) The algebraic sum of the deviations of 20 observations measured from 30 is 2. Therefore, mean of these observations is .....

(vii) The relationship between A.M., G.M. and H.M. is .....

(viii) The mean of 20 observations is 15. On checking it was found that two observations were wrongly copied as 3 and 6. If wrong observations are replaced by correct values 8 and 4, then the correct mean is .....

(ix) Median = ... Quartile.

(x) Median is the average suited for ..... classes.

(xi) A distribution with two modes is called ..... and with more than two modes is called .....

(xii) ..... is not affected by extreme observations.

Ans. *(ii)* 8 ; *(iii)* 1,748 ; *(iv)* H.M. ; *(v)* 50·1, 52 ; *(vi)* 30·1 ; *(vii)* A.M. ≥ G.M. ≥ H.M. ; *(viii)* 15·15 ; *(ix)* Second ; *(x)* Open end ; *(xi)* Bimodal, multimodal ; *(xii)* Median or mode.

VII. For the questions given below, give correct answers.

(i) The algebraic sum of the deviations of a set of n values from their arithmetic mean is

(a) n, *(b)* 0, *(c)* 1, *(d)* none of these.

(ii) The most stable measure of central tendency is

(a) the mean, *(b)* the median, *(c)* the mode, *(d)* none of these.

(iii) 10 is the mean of a set of 7 observations and 5 is the mean of a set of 3 observations. The mean of a combined set is given by

(a) 15, *(b)* 10, *(c)* 8·5, *(d)* 7·5, *(e)* none of these.

(iv) The mean of the distribution, in which the value of x are 1, 2, ..., n, the frequency of each being unity is:

(a) $n(n+1)/2$, *(b)* $n/2$, *(c)* $(n+1)/2$, *(d)* none of these.

(v) The arithmetic mean of the numbers 1, 2, 3, ..., n is

(a) $\dfrac{n(n+1)(2n+1)}{6}$, *(b)* $\dfrac{n(n+1)^2}{4}$, *(c)* $\dfrac{n(n+1)}{2}$, *(d)* none of these.

(ii) The most stable measure of central tendency

(vi) The point of intersection of the 'less than' and the 'greater than' ogive corresponds to

(a) the mean, *(b)* the median, *(c)* the geometric mean, *(d)* none of these.

(vii) When x_i and y_i are two variables ($i = 1, 2, ..., n$) with G.M.'s G_1 and G_2 respectively then the geometric mean of $\left(\dfrac{x_i}{y_i}\right)$ is

(a) $\dfrac{G_1}{G_2}$, *(b)* antilog $\left(\dfrac{G_1}{G_2}\right)$, *(c)* $n(\log G_1 - \log G_2)$;

(d) Antilog $\left(\frac{\log G_1 - \log G_2}{2n}\right)$

Ans. *(i)* (b) ; *(ii)* (a) ; *(iii)* (c) ; *(iv)* (c) ; *(v)* (d) ; *(vi)* (b) ; *(vii)* (a).

VIII. State which of the following statements are True and which are False. In case of false statements give the correct statement.

(i) The harmonic mean of n numbers is the reciprocal of the Arithmetic mean of the reciprocals of the numbers.

(ii) For the wholesale manufacturers interested in the type which is usually in demand, median is the most suitable average.

(iii) The algebraic sum of the deviations of a series of individual observations from their mean is always zero.

(iv) Geometric mean is the appropriate average when emphasis in on the rate of change rather than the amount of change.

(v) Harmonic mean becomes zero when one of the items is zero.

(vi) Mean lies between median and mode.

(vii) Cumulative frequency is not-decreasing.

(viii) Geometric mean is the arithmetic mean of harmonic mean and arithmetic mean.

(ix) Mean, median mode have the same unit.

(x) One quintal of wheat was purchased at 0·8 kg. per rupee and another quintal at 1·2 kg. per rupee. The average rate per rupee is 1kg.

(xi) One limitation of the median is that it cannot be calculated from a frequency distribution with open end classes.

(xii) The arithmetic mean of a frequency distribution is always located in the class which has the greatest number of frequencies.

(xiii) In a moderately asymmetrical distribution, the mean, median and mode are the same.

(xiv) It is really immaterial in which class an item falling at the boundary between two classes is listed.

(xv) The median is not affected by extreme items.

(xvi) The median is the point about which the sum of squared deviations is minimum.

(xvii) In construction of the frequency distribution, the selection of the class interval is arbitrary.

(xviii) Usual attendance of B.Sc. class is 35 per day. So for 100 working days total attendance is 3,500.

(xix) A car travels 100 miles at a speed of 40 m.p.h. and another 400 miles at a speed of 30 m.p.h. So the average speed for the whole journey is either 35 m.p.h. or 33 m.p.h.

(xx) In calculating the mean for grouped data, the assumption is made that the mean of the items in each class is equal to the mid-value of the class.

(xxi) The geometric mean of a group of numbers is less than the arithmetic mean in all cases, except in the special case in which the numbers are all the same.

(xxii) The geometric mean equals the antilog of the arithmetic mean of the logs of the values.

(xxiii) The median may be considered more typical than the mean because the median is not affected by the size of the extremes.

(xxiv) The Harmonic Mean of a series of fractions is the same as the reciprocal of the arithmetic mean of the series.

(xxv) In a frequency distribution the true value of mode cannot be calculated exactly.

IX. In each of the following cases, explain whether the description applies to mean, median or both.

(i) it can be calculated from a frequency distribution with open-end classes.

(ii) the values of all items are taken into consideration in the calculation.

(iii) the values of extreme items do not influence the average.

(iv) In a distribution with a single peak and moderate skewness to the right it is closer to the concentration of the distribution.

Ans. *(i)* median, *(ii)* mean, *(iii)* median, *(iv)* median.

X. Be brief in your answer :

(a) The production in an industrial unit was 10,000 units during 1981 and in 1980 the production was 25,000 units. Hence the production has declined by 150 percent. Comment.

(b) A man travels by a car for 4 days. He travelled for 10 hours each day. He drove on the first day at the rate of 45 km per hour, second day at 40 km. per hour, third day at the rate of 38 km. per hour and the fourth day at the rate of 37 km. per hour.

Which average, harmonic mean or arithmetic mean or median will give us his average speed? Why?

(c) It is seen from records that a country does not export more than 5 % of its total production. Hence export trade is not vital to the economy of that country. Is the conclusion right?

(d) A survey revealed that the children of engineers, doctors and lawyers have high intelligence quotients. It further revealed that the grandfathers of these children were also highly intelligent. Hence the inference is that intelligence is hereditary. Do you agree?

XI. Do you agree with the following interpretations made on the basis of the facts given. Explain briefly your answer.

(a) The number of deaths in military in the recent war was 10 out of 1,000 while the number of deaths in Hyderabad in the same period was 18 per 1,000. Hence it is safe to join military service than to live in the city of Hyderabad.

(b) The examination result in a college X was 70% in the year 1991. In the same year and at the same examination only 500 out of 750 students were successful in college Y. Hence the teaching standard in college X was better.

(c) The average daily production in a small-scale factory in January 1991 was 4,000 candles and 3,800 candles in February 1981. So the workers were more efficient in January.

(d) The increase in the price of a commodity was 25%. Then the price decreased by 20% and again increased by 10%. So the resultant increase in the price was $25 - 20 + 10 = 15\%$.

(e) The rate of tomato in the first week of January was 2 kg. for a rupee and in the 2nd week was 4 kg. for a rupee. Hence the average price of tomato is $\frac{1}{2}(2 + 4) = 3$ kg. for a rupee.

XII. *(a)* The mean mark of 100 students was given to be 40. It was found later that a mark 53 was read as 83. What is the corrected mean mark?

(b) The mean salary paid to 1,000 employees of an establishment was found to be Rs. 108·40. Later on, after disbursement of salary it was discovered that the salary of two employees was wrongly entered as Rs. 297 and Rs. 165. Their correct salaries were Rs. 197 and Rs. 185. Find the correct arithmetic mean.

(c) Twelve persons gambled on a certain night. Seven of them lost at an average rate of Rs. 10·50 while the remaining five gained at an average of Rs. 13·00. Is the information given above correct? If not, why?

CHAPTER THREE

Measures of Dispersion, Skewness and Kurtosis

3·1. Dispersion. Averages or the measures of central tendency give us an idea of the concentration of the observations about the central part of the distribution. If we know the average alone we cannot form a complete idea about the distribution as will be cear from the following example.

Consider the series *(i)* 7, 8, 10, 11, *(ii)* 3, 6, 9, 12, 15, *(iii)* 1, 5, 9, 13, 17. In all these cases we see that n, the number of observations is 5 and the mean is 9. If we are given that the mean of 5 observations is 9, we cannot form an idea as to whether it is the average of first series or second series or third series or of any other series of 5 observations whose sum is 45. Thus we see that the measures of central tendency are inadequate to give us a complete idea of the distribution. They must be supported and supplemented by some other measures. One such measure is *Dispersion.*

Literal meaning of dispersion is *'scatteredness'*. We study dispersion to have an idea about the homogeneity or heterogeneity of the distribution. In the above case we say that series *(i)* is more homogeneous (less dispersed) than the series *(ii)* or *(iii)* or we say that series *(iii)* is more heterogeneous (more scattered) than the series *(i)* or *(ii)*.

3·2. Characteristics for an Ideal Measure of Dispersion. The desiderata for an ideal measure of dispersion are the same as those for an ideal measure of central tendency, *viz.,*

(i) It should be rigidly defined.

(ii) It should be easy to calculate and easy to understand.

(iii) It should be based on all the observations.

(iv) It should be amenable to further mathematical treatment.

(v) It should be affected as little as possible by fluctuations of sampling.

3·3. Measures of Dispersion. The following are the measures of dispersion:

(i) Range,

(ii) Quartile deviation or Semi-interquartile range,

(iii) Mean deviation, and

(iv) Standard deviation.

3·4. Range. The range is the difference between two extreme obsertations of the distribution. If A and B are the greatest and smallest observations respectively in a distribution, then its range is $A - B$.

Range is the simplest but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to chance fluctuations, it is not at all a reliable measure of dispersion.

3·5. Quartile Deviation. Quartile deviation or semi-interquartile range

Q is given by

$$Q = \tfrac{1}{2}(Q_3 - Q_1), \qquad ...(3\cdot1)$$

where Q_1 and Q_3 are the first and third quartiles of the distribution respectively.

Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it cannot be regarded as a reliable measure.

3·6. Mean Deviation. If $x_i | f_i$, $i = 1, 2, ..., n$ is the frequency distribution, then mean deviation from the average A, (usually mean, median or mode), is given by

$$\text{Mean deviation} = \frac{1}{N} \sum_i f_i \,| x_i - A |, \quad \sum f_i = N \qquad ...(3\cdot2)$$

where $| x_i - A |$ represents the modulus or the absolute value of the deviation $(x_i - A)$, when the –ive sign is ignored.

Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations $(x_i - A)$ creates artificiality and renders it useless for further mathematical treatment.

It may be pointed out here that mean deviation is least when taken from median. (The proof is given for continuous variable in Chapter 5)

3·7. Standard Deviation and Root Mean Square Deviation. Standard deviation, usually denoted by the Greek letter small sigma (σ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution $x_i | f_i$, $i = 1, 2, ..., n$,

$$\sigma = \sqrt{\frac{1}{N} \sum_i f_i (x_i - \bar{x})^2} \qquad ...(3\cdot3)$$

where $\bar{x}$ is the arithmetic mean of the distribution and $\sum_i f_i = N$.

The step of squaring the deviations $(x_i - \bar{x})$ overcomes the drawback of ignoring the signs in mean deviation. Standard deviation is also suitable for further mathematical treatment (§ 3·7·3). Moreover of all the measures, standard deviation is affected least by fluctuations of sampling.

Thus we see that standard deviation satisfies almost all the properties laid down for an ideal measure of dispersion except for the general nature of extracting the square root which is not readily comprehensible for a non-mathematical person. It may also be pointed out that standard deviation gives greater weight to extreme values and as such has not found favour with economists or businessmen who are more interested in the results of the modal class. Taking into consideration the pros and cons and also the wide applications of standard deviation in statistical theory, we may regard standard deviation as the best and the most powerful measure of dispersion.

The square of standard deviation is called the *variance* and is given by

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 \qquad ...(3\cdot3a)$$

Root mean square deviation, denoted by 's' is given by

$$s = \sqrt{\frac{1}{N} \sum_i f_i (x_i - A)^2} \qquad ...(3\cdot4)$$

where A is any arbitrary number. s^2 is called mean square deviation.

3·7·1. Relation between σ and *s*. By definition, we have

$$s^2 = \frac{1}{N} \sum_i f_i (x_i - A)^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x} + \bar{x} - A)^2$$

$$= \frac{1}{N} \sum_i f_i \left[(x_i - \bar{x})^2 + (\bar{x} - A)^2 + 2(\bar{x} - A)(x_i - \bar{x}) \right]$$

$$= \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 + (\bar{x} - A)^2 \frac{1}{N} \sum_i f_i + 2(\bar{x} - A) \sum_i f_i (x_i - \bar{x}),$$

$(\bar{x} - A)$, being constant is taken outside the summation sign. But $\sum_i f_i (x_i - \bar{x}) = 0$, being the algebraic sum of the deviations of the given values from their mean. Thus

$$s^2 = \sigma^2 + (\bar{x} - A)^2 = \sigma^2 + d^2, \text{ where } d = \bar{x} - A$$

Obviously s^2 will be least when $d = 0$, *i.e.*, $\bar{x} = A$. Hence mean square deviation and consequently root mean square deviation is least when the deviations are taken from $A = \bar{x}$, *i.e.*, standard deviation is the least value of root mean square deviation.

The same result could be obtained alternatively as follows:
Mean square deviation is given by

$$s^2 = \frac{1}{N} \sum_i f_i (x_i - A)^2$$

It has been shown in § 2·5·1 Property 2 that $\sum_i f_i (x_i - A)^2$ is minimum when $A = \bar{x}$. Thus mean square deviation is minimum when $A = \bar{x}$ and its minimum value is

$$(s^2)\min = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \sigma^2$$

Hence variance is the minimum value of mean square deviation or standard deviation is the minimum value of root mean square deviation.

3·7·2. Different Formulae For Calculating Variance. By definition, we have

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2$$

More precisely we write it as σ_x^2, *i.e.*, variance of x. Thus

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 \qquad ...(3\cdot5)$$

If $\bar{x}$ is not a whole number but comes out to be in fractions, the calculation of σ_x^2 by using (3·5) is very cumbersome and time consuming. In order to overcome this difficulty, we shall develop different forms of the formula (3·5) which reduce the arithmetic to a great extent and are very useful for computational work. In the following sequence the summation is extended over i from 1 to n.

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i f_i (x_i^2 + \bar{x}^2 - 2\, x_i \bar{x})$$

$$= \frac{1}{N} \sum_i f_i x_i^2 + \bar{x}^2 \frac{1}{N} \sum_i f_i - 2\,\bar{x} \,.\, \frac{1}{N} \sum_i f_i x_i$$

$$= \frac{1}{N} \sum_i f_i x_i^2 + \bar{x}^2 - 2\,\bar{x}^2 = \frac{1}{N} \sum_i f_i x_i^2 - \bar{x}^2 \qquad ...(3\cdot6)$$

$$\Rightarrow \qquad \sigma_x^2 = \frac{1}{N} \sum_i f_i x_i^2 - \left(\frac{1}{N} \sum_i f_i x_i \right)^2 \qquad ...(3\cdot6a)$$

If the values of x and f are large the calculation of fx, fx^2 is quite tedious. In that case we take the deviations from any arbitrary point 'A'. Generally the point in the middle of the distribution is much convenient though the formula is true in general. We have

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i f_i (x_i - A + A - \bar{x})^2$$

$$= \frac{1}{N} \sum_i f_i (d_i + A - \bar{x})^2, \text{ where } d_i = x_i - A.$$

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i [\, d_i^2 + (A - \bar{x})^2 + 2\,(A - \bar{x})\, d_i \,]$$

$$= \frac{1}{N} \sum_i f_i d_i^2 + (A - \bar{x})^2 + 2\,(A - \bar{x}) \,.\, \frac{1}{N} \sum_i f_i d_i$$

We know that if $d_i = x_i - A$ then $\bar{x} = A + \frac{1}{N} \sum_i f_i d_i$

$$\therefore \qquad A - \bar{x} = - \frac{1}{N} \sum_i f_i d_i$$

Hence

$$\sigma_x^2 = \frac{1}{N} \sum f_i d_i^2 + \left(- \frac{1}{N} \sum_i f_i d_i \right)^2 + 2 \left(- \frac{1}{N} \sum_i f_i d_i \right) \left(\frac{1}{N} \sum_i f_i d_i \right)$$

$$= \frac{1}{N} \sum_i f_i d_i^2 - \left(\frac{1}{N} \sum_i f_i d_i \right)^2 \qquad ...(3\cdot7)$$

$$\Rightarrow \qquad \sigma_x^2 = \sigma_d^2 \qquad \text{[On comparison with (3·6a)]}$$

Hence variance and consequently standard deviation is independent of change of origin.

If we take $d_i = (x_i - A)/h$ so that $(x_i - A) = hd_i$, then

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i f_i (x_i - A + A - \bar{x})^2$$

$$= \frac{1}{N} \sum_i f_i (hd_i + A - \bar{x})^2$$

$$= h^2 \frac{1}{N} \sum_i f_i d_i^2 + (A - \bar{x})^2 + 2(A - \bar{x}) \cdot h \cdot \frac{1}{N} \sum_i f_i d_i$$

Using $\bar{x} = A + h \dfrac{\Sigma f_i d_i}{N}$, we get

$$\sigma_x^2 = h^2 \left[\frac{1}{N} \sum_i f_i d_i^2 - \left(\frac{1}{N} \sum_i f_i d_i \right)^2 \right] = h^2 \sigma_d^2, \qquad \ldots(3\cdot8)$$

which shows that *variance is not independent of change of scale.*

Aliter. If $d_i = \dfrac{x_i - A}{h}$, then

$$x_i = A + hd_i \quad \text{and} \quad \bar{x} = A + h \cdot \frac{1}{N} \sum_i f_i d_i = A + h\bar{d}$$

Obviously $\qquad x_i - \bar{x} = h(d_i - \bar{d})$

$$\therefore \qquad \sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = h^2 \cdot \frac{1}{N} \sum_i f_i (d_i - \bar{d})^2 = h^2 \sigma_d^2$$

Hence variance is independent of change of origin but not of scale.

Example 3·1. *Calculate the mean and standard deviation for the following table giving the age distribution of 542 members.*

Age in years :	20—30	30—40	40—50	50—60	60—70	70—80	80—90
No. of members :	3	61	132	153	140	51	2

Solution. Here we take $d = \dfrac{x - A}{h} = \dfrac{x - 55}{10}$

Age group	*Mid-value* (x)	*Frequncy* (f)	$d = \dfrac{x - 55}{10}$	fd	fd^2
20 — 30	25	3	−3	−9	27
30 — 40	35	61	−2	−122	244
40 — 50	45	132	−1	−132	132
50 — 60	55	153	0	0	0
60 — 70	65	140	1	140	140
70 — 80	75	51	2	102	204
80 — 90	85	2	3	6	18
		$N = \Sigma f = 542$		$\Sigma fd = -15$	$\Sigma fd^2 = 765$

$$\bar{x} = A + h \frac{\Sigma fd}{N} = 55 + \frac{10 \times (-15)}{542} = 55 - 0\cdot28 = 54\cdot72 \text{ years.}$$

$$\sigma^2 = h^2 \left[\frac{1}{N} \Sigma fd^2 - \left(\frac{1}{N} \Sigma fd \right)^2 \right] = 100 \left[\frac{765}{542} - (0\cdot28)^2 \right]$$

$= 100 \times 1{\cdot}333 = 133{\cdot}3$

$\therefore$ σ (standard deviation) = 11·55 years

Example 3·2. *Prove that for any discrete distribution standard deviation is not less than mean deviation from mean.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

Solution. Let $x_i \mid f_i$, $i = 1, 2, 3, ..., n$ be any discrete distribution. Then we have to prove that

$$\text{S.D.} \nless \text{Mean deviation from mean}$$

$$\Rightarrow \quad (\text{S.D.})^2 \nless (\text{Mean deviation from mean})^2$$

$$\Rightarrow \quad (\text{S.D.})^2 \geq (\text{M. D. from mean})^2$$

$$\Rightarrow \quad \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2 \geq \left(\frac{1}{N}\sum_{i=1}^{n} f_i |x_i - \bar{x}|\right)^2$$

If we put $|x_i - \bar{x}| = z_i$, then we have to prove that

$$\frac{1}{N}\sum_{i=1}^{n} f_i z_i^2 \geq \left(\frac{1}{N}\sum_{i=1}^{n} f_i z_i\right)^2$$

i.e., $$\frac{1}{N}\sum_{i=1}^{n} f_i z_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i z_i\right)^2 \geq 0$$

i.e., $$\frac{1}{N}\sum_{i=1}^{n} f_i (z_i - \bar{z})^2 \geq 0$$

i.e., $$\sigma_z^2 \geq 0,$$

which is always true. Hence the result.

Example 3·3. *Find the mean deviation from the mean and standard deviation of A.P. $a, a + d, a + 2d, ..., a + 2nd$ and verify that the latter is greater than the former.* **[Delhi Univ. B.Sc. (Stat. Hons.), 1990]**

Solution. We know that the mean of a series in A.P. is the mean of its first and last term. Hence the mean of the given series is

$$\bar{x} = \frac{1}{2}(a + a + 2nd) = a + nd$$

x	$\lvert x - \bar{x} \rvert$	$(x - \bar{x})^2$
a	nd	n^2d^2
$a + d$	$(n-1)d$	$(n-1)^2 d^2$
$a + 2d$	$(n-2)d$	$(n-2)^2 d^2$
$\vdots$	$\vdots$	$\vdots$
$a + (n-2)d$	$2d$	$2^2 . d^2$
$a + (n-1)d$	d	$1^2 . d^2$
$a + nd$	0	0
$a + (n+1)d$	d	$1^2 . d^2$
$a + (n+2)d$	$2d$	$2^2 . d^2$

⋮	⋮	⋮
$a + (2n - 2)\, d$	$(n - 2)\, d$	$(n - 2)^2 d^2$
$a + (2n - 2)\, d$	$(n - 1)\, d$	$(n - 2)^2 d^2$
$a + 2nd$	nd	$n^2 d^2$

$$\text{Mean deviation from mean} = \frac{1}{2n+1} \Sigma \,|\, x - \bar{x} \,|$$

$$= \frac{1}{2n+1} 2.d\,(1 + 2 + 3 + \dots + n)$$

$$= \frac{n\,(n+1)d}{(2n+1)}$$

$$\sigma^2 = \frac{1}{2n+1} \Sigma\,(x - \bar{x})^2 = \frac{1}{2n+1} 2.d^2\,(1^2 + 2^2 + 3^2 + \dots + n^2)$$

$$= \frac{1}{2n+1} 2d^2 \cdot \frac{n\,(n+1)\,(2n+1)}{6} = \frac{n\,(n+1)d^2}{3}$$

Hence standard deviation

$$\sigma = \sqrt{\frac{n\,(n+1)}{3}} \times d$$

Verification.

$$\text{S.D.} > \text{M.D. from mean}$$

if $\quad (\text{S.D.})^2 > (\text{M.D. from mean})^2$

i.e., if $\quad \dfrac{n\,(n+1)d^2}{3} > \left(\dfrac{n\,(n+1)d}{2n+1} \right)^2$

or if $\quad (2n+1)^2 > 3n\,(n+1)$

or if $\quad n^2 + n + 1 > 0$

or if $\quad (n + \frac{1}{2})^2 + \frac{3}{4} > 0$

which is always ture.

Example 3·4. *Show that in a discrete series if deviations are small compared with mean M so that $(x/M)^3$ and higher powers of (x/M) are neglected, we have*

$$\text{(i) } G = M \left(1 - \frac{1}{2} \cdot \frac{\sigma^2}{M^2} \right),$$

$$\text{(ii) } M^2 - G^2 = \sigma^2, \text{ and (iii) } H = M \left(1 - \frac{\sigma^2}{M^2} \right),$$

where M is the arithemetic mean, G, the grometric mean, H, the harmonic mean and σ is the standard deviation of the distribution.

Solution. Let $X_i \,|\, f_i$, $i = 1, 2, \dots, n$ be the given frequency distribution. Then we are given that $x_i = X_i - M$, *i.e.*, $X_i = x_i + M$ where M is the mean of the distribution. We have

$$\sum_i f_i x_i = \sum_i f_i (X_i - M) = 0, \quad ...(1)$$

being the algebraic sum of the deviations of the given values from their mean. Also

$$\sum_i f_i x_i^2 = \sum_i f_i (X_i - M)^2 = \sigma^2 \quad ...(2)$$

(i) By definition, we have

$$G = (X_1^{f_1} . X_2^{f_2} \dots X_n^{f_n})^{1/N}, \text{ where } N = \Sigma f_i$$

$$\log G = \frac{1}{N} \sum_i f_i \log X_i = \frac{1}{N} \sum_i f_i \log (x_i + M)$$

$$= \frac{1}{N} \sum_i \left\{ f_i \log \left[M \left(1 + \frac{x_i}{M} \right) \right] \right\}$$

$$= \frac{1}{N} \sum_i \left\{ f_i \left[\log M + \log \left(1 + \frac{x_i}{M} \right) \right] \right\}$$

$$= \log M + \frac{1}{N} \sum_i f_i \log \left(1 + \frac{x_i}{M} \right)$$

$$= \log M + \frac{1}{N} \sum_i f_i \left[\frac{x_i}{M} - \frac{1}{2} \frac{x_i^2}{M^2} + \frac{1}{3} \left(\frac{x_i}{M} \right)^3 + \dots \right],$$

the expansion of $\log \left(1 + \frac{x_i}{M} \right)$ in ascending powers of (x_i/M) being valid since $| x_i / M | < 1$. Neglecting $(x_i/M)^3$ and higher powers of (x_i/M), we get

$$\log G = \log M + \frac{1}{NM} \sum_i f_i x_i - \frac{1}{2M^2} \cdot \frac{1}{N} \sum_i f_i x_i^2$$

$$= \log M - \frac{\sigma^2}{2M^2}, \qquad \text{[On using (1) and (2)]}$$

$$= \log \left(M e^{-\sigma^2/2M^2} \right)$$

$$\Rightarrow \qquad G = M e^{-\sigma^2/2M^2} = M \left(1 - \frac{\sigma^2}{2M^2} \right),$$

neglecting higher powers.

Hence $$G = M \left(1 - \frac{1}{2} \cdot \frac{\sigma^2}{M^2} \right). \quad ...(3)$$

(ii) Squaring both sides in (3), we get

$$G^2 = M^2 \left(1 - \frac{1}{2} \cdot \frac{\sigma^2}{M^2} \right)^2 = M^2 \left(1 - \frac{\sigma^2}{M^2} \right) = M^2 - \sigma^2,$$

neglecting $(\sigma/M)^4$.

$$\therefore \qquad M^2 - G^2 = \sigma^2 \quad ...(4)$$

(iii) By definition, harmonic mean H is given by

$$\frac{1}{H} = \frac{1}{N} \sum_i (f_i/X_i) = \frac{1}{N} \sum_i [f_i/(x_i + M)]$$

$$= \frac{1}{MN} \sum_i \frac{f_i}{[1 + (x_i/M)]} = \frac{1}{MN} \sum_i f_i \left(1 + \frac{x_i}{M}\right)^{-1}$$

Since $\left|\frac{x_i}{M}\right| < 1$, the expansion of $\left(1 + \frac{x_i}{M}\right)^{-1}$ in ascending powers of (x_i/M) is valid. Neglecting $(x_i/M)^3$ and higher powers of (x_i/M), we get

$$\frac{1}{H} = \frac{1}{MN} \sum_i f_i \left(1 - \frac{x_i}{M} + \frac{x_i^2}{M^2}\right)$$

$$= \frac{1}{M}\left(\frac{1}{N} \sum_i f_i - \frac{1}{MN} \sum_i f_i x_i + \frac{1}{M^2}\frac{1}{N} \sum_i f_i x_i^2\right)$$

$$= \frac{1}{M}\left(1 + \frac{\sigma^2}{M^2}\right) \qquad \text{[On using (1) and (2)]}$$

$$\therefore \qquad H = M\left(1 + \frac{\sigma^2}{M^2}\right)^{-1} = M\left(1 - \frac{\sigma^2}{M^2}\right),$$

higher powers being neglected.

Hence $$H = M\left(1 - \frac{\sigma^2}{M^2}\right) \qquad ...(5)$$

Example 3·5. *For a group of 200 candidates, the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered that the scores 43 and 35 were misread as 34 and 53 respectively. Find the corrected mean and standard deviation corresponding to the corrected figures.*

Solution. Let x be the given variable. We are given $n = 200$, $\bar{x} = 40$ and $\sigma = 15$

Now $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \Rightarrow \quad \sum_i x_i = n\bar{x} = 200 \times 40 = 8000$$

Also $$\sigma^2 = \frac{1}{n} \sum_i x_i^2 - \bar{x}^2$$

$$\therefore \qquad \sum_i x_i^2 = n(\sigma^2 + \bar{x}^2) = 200\,(225 + 1600) = 365000$$

$$\text{Corrected } \sum_i x_i = 8000 - 34 - 53 + 43 + 35 = 7991$$

and $$\text{Corrected } \sum_i x_i^2 = 365000 - (34)^2 - (53)^2 + (43)^2 + (35)^2 = 364109$$

Hence, $$\text{Corrected mean} = \frac{7991}{200} = 39{\cdot}955$$

Corrected $\sigma^2 = \dfrac{364109}{200} - (39{\cdot}955)^2 = 1820{\cdot}54 - 1596{\cdot}40 = 224{\cdot}14$

$\therefore$ Corrected standard deviation $= 14{\cdot}97$

3·7·3. Theorem. *(Variance of the combined series). If n_1, n_2 are the sizes; $\bar{x}_1, \bar{x}_2$ the means, and σ_1, σ_2 the standard deviations of two series, then the standard deviation σ of the combined series of size $n_1 + n_2$ is given by*

$$\sigma^2 = \frac{1}{n_1 + n_2}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)\right] \qquad \ldots(3{\cdot}9)$$

where $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$

and $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$, is the mean of the combined series.

Proof. Let x_{1i}; $i = 1, 2, \ldots, n_1$ and x_{2j}; $j = 1, 2, \ldots, n_2$, be the two series then

$$\left.\begin{aligned}\bar{x}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}\\ \bar{x}_2 &= \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}\end{aligned}\right\}\ldots(*) \quad \text{and} \quad \left.\begin{aligned}\sigma_1^2 &= \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2\\ \sigma_2^2 &= \frac{1}{n_2}\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2\end{aligned}\right\} \qquad \ldots(**)$$

The mean $\bar{x}$ of the combined series is given by

$$\bar{x} = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j}\right] = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \qquad \text{[From (*)]}$$

The variance σ^2 of the combined series is given by

$$\sigma^2 = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x})^2\right] \qquad \ldots(3{\cdot}10)$$

Now

$$\begin{aligned}\sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 &= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1 + \bar{x}_1 - \bar{x})^2\\ &= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + n_1(\bar{x}_1 - \bar{x})^2 + 2(\bar{x}_1 - \bar{x})\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) \quad \ldots(3{\cdot}10a)\end{aligned}$$

But $\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) = 0$, being the algebraic sum of the deviations the values of first series from their mean. Hence from (3·10*a*), on using (**), we get

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 = n_1\sigma_1^2 + n_1(\bar{x}_1 - \bar{x})^2 = n_1\sigma_1^2 + n_1 d_1^2 \qquad \ldots(3{\cdot}10b)$$

where $d_1 = \bar{x}_1 - \bar{x}$.

Similarly, we get

$$\sum_{j=1}^{n_2} (x_{2j} - \bar{x})^2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2 + \bar{x}_2 - \bar{x})^2$$

$$= \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2 + n_2 (\bar{x}_2 - \bar{x})^2 = n_2 \sigma_2^2 + n_2 d_2^2 \qquad \text{...(3·10}c\text{)}$$

where $d_2 = \bar{x}_2 - \bar{x}$.

Substituting from (3·10*b*) and (3·10*c*) in (3·10), we get the required formula

$$\sigma^2 = \frac{1}{n_1 + n_2} \left[n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2) \right]$$

This formula can be simplified still further. We have

$$d_1 = \bar{x}_1 - \bar{x} = \bar{x}_1 - \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{n_2 (\bar{x}_1 - \bar{x}_2)}{n_1 + n_2}$$

$$d_2 = \bar{x}_2 - \bar{x} = \bar{x}_2 - \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{n_1 (\bar{x}_2 - \bar{x}_1)}{n_1 + n_2}$$

Hence

$$\sigma^2 = \frac{1}{n_1 + n_2} \left[n_1 \sigma_1^2 + n_2 \sigma_2^2 + \left\{ \frac{n_1 n_2^2 (\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2} + \frac{n_2 n_1^2 (\bar{x}_2 - \bar{x}_1)^2}{(n_1 + n_2)^2} \right\} \right]$$

$$= \frac{1}{n_1 + n_2} \left[n_1 \sigma_1^2 + n_2 \sigma_2^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2 \right] \qquad \text{...(3·11)}$$

Remark. The formula (3·9) can be easily generalised to the case of more than two series. If n_i, $\bar{x}_i$ and σ_i, $i = 1, 2, ..., k$ are the sizes, means and standard deviations respectively of k-component series then the standard deviation σ of the combined series of size $\sum_{i=1}^{k} n_i$ is given by

$$\sigma^2 = \frac{1}{n_1 + n_2 + \ldots + n_k} \left[n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2) + \ldots + n_k (\sigma_k^2 + d_k^2) \right] \qquad \text{...(3·12)}$$

where $d_i = \bar{x}_i - \bar{x}$; $i = 1, 2, ..., k$

and $\bar{x} = \dfrac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \ldots + n_k \bar{x}_k}{n_1 + n_2 + \ldots + n_k}$

Example 3·6. *The first of the two samples has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15·6 and standard deviation $\sqrt{13 \cdot 44}$, find the standard deviation of the second group.*

Solution. Here we are given

$$n_1 = 100, \ \bar{x}_1 = 15 \ \text{and} \ \sigma_1 = 3$$
$$n = n_1 + n_2 = 250, \ \bar{x} = 15 \cdot 6, \ \text{and} \ \sigma = \sqrt{13 \cdot 44}$$

We want σ_2.

Obviously $n_2 = 250 - 100 = 150$. We have

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \Rightarrow 15 \cdot 6 = \frac{100 \times 15 + 150 \times \bar{x}_2}{250}$$

$\Rightarrow \quad 150\, \bar{x}_2 = 250 \times 15 \cdot 6 - 1500 = 2400$

$\therefore \quad \bar{x}_2 = \dfrac{2400}{150} = 16$

Hence $\quad d_1 = \bar{x}_1 - \bar{x} = 15 - 15 \cdot 6 = -0 \cdot 6$

and $\quad d_2 = \bar{x}_2 - \bar{x} = 16 - 15 \cdot 6 = 0 \cdot 4$

The variance σ^2 of the combined group is given by the formula :

$$(n_1 + n_2)\sigma^2 = n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)$$

$\Rightarrow \quad 250 \times 13 \cdot 44 = 100\,(9 + 0 \cdot 36) + 150\,(\sigma_2^2 + 0 \cdot 16)$

$\therefore \quad 150\, \sigma_2^2 = 250 \times 13 \cdot 44 - 100 \times 9 \cdot 36 - 150 \times 0 \cdot 16$

$\quad = 3360 - 936 - 24 = 2400$

$\therefore \quad \sigma_2^2 = \dfrac{2400}{150} = 16$

Hence $\quad \sigma_2 = \sqrt{16} = 4$

3·8. Co-efficient of Dispersion. Whenever we want to compare the variability of the two series which differ widely in their averages or which are measured in different units, we do not merely calculate the measures of dispersion but we calculate the co-efficients of dispersion which are pure numbers independent of the units of measurement. The co-efficients of dispersion (C.D.) based on different measures of dispersion are as follows :

1. C.D. based upon range $= \dfrac{A - B}{A + B}$, where A and B are the greatest and the smallest items in the series.

2. Based upon quartile deviation :

$$\text{C.D.} = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3. Based upon mean deviation :

$$\text{C.D.} = \frac{\text{Mean deviation}}{\text{Average from which it is calculated}}$$

4. Based upon standard deviation :

$$\text{C.D.} = \frac{\text{S.D.}}{\text{Mean}} = \frac{\sigma}{\bar{x}}$$

3·8·1. Co-efficient of Variation. 100 times the co-efficient of desperison based upon standard deviation is called co-efficient of variation (C.V.),

$$\text{C.V.} = 100 \times \frac{\sigma}{\bar{x}} \qquad \ldots(3 \cdot 13)$$

According to Professor Karl Pearson who suggested this measure, C.V. *is the percentage variation in the mean, standard deviation being considered as the total variation in the mean.*

For comparing the variability of two series, we calculate the co-efficient of variations for each series. The series having greater C.V. is said to be more variable than the other and the series having lesser C.V. is said to be more consistent (or homogenous) than the other.

Example 3·7. *An analysis of monthly wages paid to the workers of two firms A and B belonging to the same industry gives the following results :*

	Firm A	Firm B
Number of workers	500	600
Average monthly wage	Rs. 186·00	Rs. 175·00
Variance of distribution of wages	81	100

(i) *Which firm, A or B, has a larger wage bill ?*

(ii) *In which firm, A or B, is there greater variability in individual wages?*

(iii) *Calculate (a) the average monthly wage, and (b) the variance of the distribution of wages, of all the workers in the firms A and B taken together.*

Solution.

(i) **Firm A :**

No. of wage-earners (say) $n_1 = 500$
Average monthly wages (say) $\bar{x}_1 = \text{Rs.}186$

$$\text{Average monthly wage} = \frac{\text{Total wages paid}}{\text{No. of workers}}$$

Hence total wages paid to the workers $= n_1\bar{x}_1 = 500 \times 186 = \text{Rs. } 93{,}000$

Firm B

No. of wage-earners (say) $n_2 = 600$
Average monthly wages (say) $\bar{x}_2 = \text{Rs.}175$

$\therefore$ Total wages paid to the workers $= n_2\bar{x}_2 = 600 \times 175 = \text{Rs. } 1{,}05{,}000$

Thus we see that the firm *B* has larger wage bill.

(ii) Variance of distribution of wages in firm *A* (say) $\sigma_1^2 = 81$
Variance of distribution of wages in firm *B* (say) $\sigma_2^2 = 100$

$$\text{C.V. of distribution of wages for firm } A = 100 \times \frac{\sigma_1}{\bar{x}_1} = \frac{100 \times 9}{186} = 4{\cdot}84$$

$$\text{C.V. of distribution of wages for firm } B = 100 \times \frac{\sigma_2}{\bar{x}_2} = \frac{100 \times 10}{175} = 5{\cdot}71$$

Since C.V. for firm *B* is greater than C.V. for firm *A*, firm *B* has greater variability in individual wages.

(iii) (a) The average monthly wages (say) $\bar{x}$, of all the workers in the two firms A and *B* taken together is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{500 \times 186 + 600 \times 175}{500 + 600} = \frac{198000}{1100} = \text{R}\ \ 180$$

(b) The combined variance σ^2 is given by the formula:

$$\sigma^2 = \frac{1}{n_1 + n_2}\left[n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2) \right]$$

where $d_1 = \bar{x}_1 - \bar{x}$ and $d_2 = \bar{x}_2 - \bar{x}$

Here $d_1 = 186 - 180 = 6$ and $d_2 = 175 - 180 = -5$

Hence $\sigma^2 = \dfrac{500(81+36)+600(100+25)}{500+600} = \dfrac{133500}{1100} = 121 \cdot 36$

EXERCISE 3 (a)

1. (a) Explain with suitable examples the term 'dispersion.' State the relative and absolute measures of dispersion and describe the merits and demerits of standard deviation.

(b) Explain the main difference between mean deviation and standard deviation. Show that standard deviation is independent of change of origin and scale.

(c) Distinguish between absolute and relative measures of dispersion.

2. *(a)* Explain the graphical method of obtaining median and quartile deviation. **(Calicut Univ.B.Sc, .April 1989)**

(b) Compute quartile deviation graphically for the following data :

Marks :	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 & over
Number of students :	5	20	14	10	8	5

3. *(a)* Show that for raw data mean deviation is minimum when measured from the median.

(b) Compute a suitable measure of dispersion for the following grouped frequency distribution giving reasons :

Classes	*Frequency*
Less than 20	30
20 – 30	20
30 – 40	15
40 – 50	10
50 – 60	5

(c) Age distribution of hundred life insurance policyholders is as follows:

Age as on nearest birthday	*Number*
17 – 19·5	9
20 – 25·5	16
26 – 35·5	12
36 – 40·5	26
41 – 50·5	14
51 – 55·5	12
56 – 60·5	6
61 – 70·5	5

Calculate mean deviation from median age.

Ans. Median = 38·25, M.D.=10·605

4. Prove that the mean deviation about the mean $\bar{x}$ of the variate x, the frequency of whose ith size x_i is f_i is given by

$$\frac{2}{N}\left[\bar{x}\sum_{x_i<\bar{x}} f_i - \sum_{x_i<\bar{x}} f_i x_i\right].$$

Hint. Mean deviation about mean

$$= \frac{1}{N}\left[\sum_{x_i<\bar{x}} f_i(\bar{x}-x_i) + \sum_{x_i>\bar{x}} f_i(x_i-\bar{x})\right]$$

$$= \frac{1}{N}\left[-\sum_{x_i<\bar{x}} f_i(x_i-\bar{x}) + \sum_{x_i>\bar{x}} f_i(x_i-\bar{x})\right]$$

Since $\{\Sigma f_i(x_i-\bar{x}) = 0,$

$$\sum_{x_i>\bar{x}} f_i(x_i-\bar{x}) + \sum_{x_i<\bar{x}} f_i(x_i-\bar{x}) = 0$$

$$\therefore\ M.D. = \frac{1}{N}\left(-\sum_{x_i<\bar{x}} f_i(x_i-\bar{x}) - \sum_{x_i<\bar{x}} f_i(x_i-\bar{x})\right) = -\frac{2}{N}\left(\sum_{x_i<\bar{x}} f_i(x_i-\bar{x})\right)$$

5. What is standard deviation ? Explain its superiority over other measures of dispersion.

6. Calculate the mean and standard deviation of the following distribution:

x :	2.5 — 7.5	7.5 — 12.5	12.5 — 17.5	17.5 — 22.5	
f :	12	28	65	121	
x :	22.5 — 27.5	27.5 — 32.5	32.5 — 37.5	37.5 — 42.5	42 5 — 47.5
f :	175	198	176	120	66
x :	47.5 — 52.5	52.5 — 57.5	57.5 — 62.5		
f :	27	9	3		

Ans. Mean = 30.005, Standard Deviation = 0.01

7. Explain clearly the ideas implied in using arbitrary working orgin, and scale for the calculation of the arithmetic mean and standard deviation of a frequency distribution. The values of the arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from the analysis in the above manner are40.604 lb. and 7.92 lb. respectively.

x :	−3	−2	−1	0	1	2	3	4	Total
f :	3	15	45	57	50	36	25	9	240

Determine the actual class intervals.

8. *(a)*. The arithmetic mean and variance of a set of 10 figures are known to be 17 and 33 respectively. Of the 10 figures, one figure (*i.e.*, 26) was subsequently found inacurate, and was weeded out. What is the resulting *(a)* arithmetic mean and *(b)* standard deviation. **(M.S. Baroda U. B.Sc. 1993)**

(b) The mean and standard deviation of 20 items is found to be 10 and 2 respectively. At the time of checking it was found that one item 8 was incorrect. Calculate the mean and standard deviation if

(i) the wrong item is omitted, and
(ii) it is replaced by 12.

(c) For a frequency distribution of marks in Statistics of 200 candidates (grouped in intervals 0–5, 5–10,..., etc.), the mean and standard deviation were found to be 40 and 15 respectively. Later it was discovered that the score 43 was misread as 53 in obtaining the frequency distribution. Find the corrected mean and standard deviation corresponding to the corrected frequency distribution.

Ans. Mean = 39·95, S.D. = 14·974.

9. *(a)* Complete a table showing the frequencies with which words of different numbers of letters occur in the extract reproduced below (omitting punctuation marks) treating as the variable the number of letters in each word, and obtain the mean, median and co-efficient of variation of the distribution :

"Her eyes were blue : blue as autumn distance-blue as the blue we see, between the retreating mouldings of hills and woody slopes on a sunny September morning : a misty and shady blue, that had no beginning or surface, and was looked into rather than at. "

Ans. Mean = 4·35, Median = 4, σ = 2·23 and C.V. = 51·26

(b) Treating the number of letters in each word in the following passage as the variable x, prepare the frequency distribution table and obtain its mean, median, mode and variance.

"The reliability of data must always be examined before any attempt is made to base conclusions upon them. This is true of all data, but particularly so of numerical data, which do not carry their quality written large on them. It is a waste of time to apply the refined theoretical methods of Statistics to data which are suspect from the beginning. "

Ans. Mean = 4·565, Median = 4, Mode = 3, S.D. = 2·673.

10. The mean of 5 observations is 4·4 and variance is 8·24. It three of the five observation are 1, 2 and 6, find the other two.

11. *(a)* Scores of two golfers for 24 rounds were as follows :

Golfer A : 74, 75, 78, 72, 77, 79, 78, 81, 76, 72, 72, 77, 74, 70, 78, 79, 80, 81, 74, 80, 75, 71, 73.

Golfer B : 86, 84, 80, 88, 89, 85, 86, 82, 82, 79, 86, 80, 82, 76, 86, 89, 87, 83, 80, 88, 86, 81, 81, 87

Find which golfer may be considered to be a more consistent player ?

Ans. Golfer B is more consistent player.

(b) The sum and sum of squares corresponding to length X (in cms.) and weight Y (in gms.) of 50 tapioca tubers are given below :

$\Sigma X = 212, \quad \Sigma X^2 = 902·8$

$\Sigma Y = 261, \quad \Sigma Y^2 = 1457·6$

Which is more varying, the length or weight.

12. *(a)* Lives of two models of refrigerators turned in for new models in a recent survey are

Life (No. of years)	*Model A*	*Model B*
0 – 2	5	2
2 – 4	16	7

4 – 6	13	12
6 – 8	7	19
8 – 10	5	9
10 – 12	4	1

What is the average life of each model of these refrigerators ? Which model shows more uniformity ?

Ans. C.V. (Model A)=54·9%, C.V. (Model B)=3·62%

(*b*) Goals scored by two teams A and B in a football season were as follows :

No. of goals scored in a match	*No. of matches* A	B
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

(Sri Venketeswara U. B.Sc. Sept. 1992)

Find out which team is more consistent.

Ans. Team A : C.V.= 122·0, Team B : C.V. = 108·3 .

(c) An analysis of monthly wages paid to the workers in two firms, A and B belonging to the same industry, gave the following results :

	Firm A	*Firm B*
No. of wage-earners	986	548
Average monthly wages	Rs. 52·5	Rs. 47·5
Variance of distribution of wages	100	121

(*i*) Which firm, A or B, pays out larger amount as monthly wages ?

(*ii*) In which firm A or B, is there greater variability in individual wages?

(*iii*) What are the measures of average monthly wages and the variability in individual wages, of all the workers in the two firms, A and B taken together.

Ans. (*i*) Firm B pays a larger amount as monthly wages.
(*ii*) There is greater variability in individual wages in firm B.
(*iii*) Combined arithmetic mean = Rs.49·87.
Combined standard deviation = Rs.10·82.

14. (*a*) The following data give the arithmetic averages and standard deviations of three sub-groups. Calculate the arithmetic average and standard deviation of the whole group.

Sub-group	*No. of men*	*Average wages (Rs.)*	*Standard deviation (Rs.)*
A	50	61·0	8·0
B	100	70·0	9·0
C	120	80·5	10·0

Ans. Combined Mean = 73, Combined S.D.=11·9 .

(*b*) Find the missing information from the following data :

	Group I	Group II	Group III	Combined
Number	50	?	90	200
Standard Deviation	6	7	?	7·746
Mean	113	?	115	116

Ans. $n_2 = 60$, $\bar{x}_2 = 120$ and $\sigma_3 = 8$

15. A collar manufacturer is considering the production of a new style collar to attract young men. The following statistics of neck circumference are avail based on the measurement of a typical group of students :

Mid-value in inches	12·5	13·0	13·5	14·0	14·5	15·0	15·5	16·0
No. of students :	4	19	30	63	66	29	18	1

Compute the mean and standard deviation and use the criterion $\bar{x} \pm$ obtain the largest and smallest size of collar he should make in order to meet needs of practically all his customers bearing in mind that the collars are worn on average 3/4 inch larger than neck size. **(Nagpur Univ. B.Sc., 1992)**

Ans. Mean = 14·232, S.D.=0·72, largest size = 17·14", smallest size = 12·83"

16. *(a)* A frequency distribution is divided into two parts. The mean and standard deviation of the first part are m_1 and s_1 and those of the second part are m_2 and s_2 respectively. Obtain the mean and standard deviation for the combined distribution. **[Delhi Univ. B.Sc.(Stat.Hons.), 1986]**

(b) The means of two samples of size 50 and 100 respectively are 54·1 and 50·3 and the standard deviations are 8 and 7. Obtain the mean and standard deviation of the sample of size 150 obtained by combining the two samples.

Ans. Combined mean = 51·57. Combined S.D. = 7·5 approx.

(c) A distribution consists of three components with frequencies 200, 250 and 300 having means 25, 10 and 15 and standard deviations 3, 4 and 5 respectively.

Show that the mean of the combined group is 16 and its standard deviation 7.2 approximately. **(Bangalore Univ. B.Sc. 1992)**

17. In a certain test for which the pass marks is 30, the distribution of marks of passing candidates classified by sex (boys and girls) were as given below :

Marks	*Frequency*	
	Boys	*Girls*
30-34	5	15
35-39	10	20
40-44	15	30
45-49	30	20
50-54	5	5
55-59	5	-
Total	70	90

The overall means and standard deviation of marks for boys including the 30 failed were 38 and 10. The corresponding figures for girls including the 10 failed were 35 and 9.

(i) Find the mean and standard deviation of marks obtained by the 30 boys who failed in the test.

(ii) The moderation committee argued that percentage of passes among girls is higher because the girls are very studious and if the intention is to pass those who are really intelligent, a higher pass marks should be used for girls. Wiothout quetioning the propriety of this argument, suggest what the pass mark should be which would allow only 70% of the girls to pass.

(iii) The prize committee decided to award prizes to the best 40 candidates (irrespective of sex) judged on the basis of marks obtained in the test. Estimate the number of girls who would receive prizes.

Ans. (i) $\bar{x} = 22{\cdot}83$, $\sigma_2 = 8{\cdot}27$ (ii) 39 (iii) 15

18. Find the mean and variance of first n-natural numbers.

(Agra Univ. B.Sc., 1993)

Ans. $\bar{x} = \dfrac{n+1}{2}$, $\sigma_2 = \dfrac{n^2-1}{12}$

19. In a frequency distribution, the n intervals are 0 to 1, 1 to 2, ..., $(n-1)$ to n with equal frequencies. Find the mean deviation and variance.

20. If the mean and standard deviation of a variable x are m and σ respectively, obtain the mean and standard deviation of $(ax + b)/c$, where a, b and c are constants.

Ans. $\bar{u} = \dfrac{1}{c}(a\bar{x} + b)$, $\sigma_u = \left|\dfrac{a}{c}\right| \sigma$

21. In a series of measurements we obtain m_1 values of magnitude x_1, m_2 values of magnitude x_2, and so on. If $\bar{x}$ is the mean value of all the measurements, prove that the standard deviation is

$$\sqrt{\frac{\Sigma m_r (k - x_r)^2}{\Sigma m_r} - \delta^2}$$

where $\bar{x} = k + \delta$ and k is any constant. **Delhi Univ. B.Sc. (Stat. Hons.), 1992**

22. (a) Show that in a discrete series if deviations are small compared with mean M so that $(x/M)^2$ and higher powers of (x/M) are neglected, prove that

(i) $MH = G^2$ *(II)* $M - 2G + H = 0$,

where G is geometric mean and H is harmonic mean.

(b) The mean and standard deviation of a variable x are m and σ respectively. If the deviations are small compared with the value of the mean, show that

(i) Mean $(\sqrt{x}) = \sqrt{m}\left(1 - \dfrac{\sigma^2}{8m^2}\right)$

(ii) Mean $\left(\dfrac{1}{\sqrt{x}}\right) = \dfrac{1}{\sqrt{m}}\left(1 + \dfrac{3\sigma^2}{8m^2}\right)$ approximately.

(M.S. Baroda U. B.Sc. 1993)

(*c*) If the deviation $X_i = x_i - M$ is very small in comparison with mean M and $(X_i/M)^2$ and higher powers of (X_i/M) are neglected, prove that

$$V = \sqrt{\frac{2(M-G)}{M}}$$

where G is the geometric mean of the values $x_1, x_2, \ldots, x_n$ and V is the coefficient of dispersion (σ/M). **(Lucknow Univ. B.Sc., 1993)**

23. From a sample of observations the arithmetic mean and variance are calculated. It is then found that one of the values, x_1, is in error and should be replaced by x_1'. Show that the adjustment to the variance to correct this error is

$$\frac{1}{n}(x_1' - x_1)\left(x_1' + x_1 - \frac{x_1' - x_1 + 2T}{n}\right)$$

where T is the total of the original results.

(Meerut Univ. B.Sc., 1992; Delhi Univ. B.Sc. (Stat. Hons.), 1989, 1985]

Hint. $\sigma^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$

$$= \frac{1}{n}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right) - \frac{T^2}{n^2}$$

where $T = x_1 + x_2 + \ldots + x_n$.

Let σ_1^2 be the corrected variance. Then

$$\sigma_1^2 = \frac{1}{n}\left\{x_1'^2 + x_2^2 + \ldots + x_n^2\right\} - \left\{\frac{T - x_1 + x_1'}{n}\right\}^2$$

Adjustment to the variance to correct the error is :

$$\sigma_1^2 - \sigma^2 = \frac{1}{n}\left\{x_1'^2 - x_1^2\right\} - \frac{1}{n^2}\left\{\left(T - x_1 + x_1'\right)^2 - T^2\right\}$$

$$= \frac{1}{n}\left\{x_1' + x_1\right\}\left\{x_1' - x_1\right\} - \frac{1}{n^2}\left\{\left(x_1' - x_1\right) \times (2T - x_1 + x_1')\right\}$$

24. Show that, if the variable takes the values $0, 1, 2, \ldots, n$ with frequencies proportional to the binomial coefficients $^nC_0, {}^nC_1, {}^nC_2, \ldots, {}^nC_n$ respectively then the mean of the distribution is $(n/2)$, the mean square deviation about $x = 0$ is $n(n+1)/4$ and the variance is $n/4$

[Delhi Univ. B.Sc. (Stat. Hons.), 1991]

Hint. $N = \sum f = {}^nC_0 + {}^nC_1 + {}^nC_2 + \ldots + {}^nC_n = (1+1)^n = 2^n$

$$\sum fx = 0.{}^nC_0 + 1.{}^nC_1 + 2.{}^nC_2 + 3.{}^nC_3 + \ldots + n.{}^nC_n$$

$$= n\left\{1 + (n-1) + \frac{(n-1)(n-2)}{2!} + \ldots + 1\right\}$$

$$= n(1+1)^{n-1} = n \cdot 2^{(n-1)}$$

Hence mean $(\bar{x}) = \dfrac{1}{N}\sum fx = \dfrac{n \cdot 2^{(n-1)}}{2^n} = \dfrac{n}{2}$

The mean square deviation s^2, (say), about the point $x = 0$ is given by

$$s^2 = \frac{1}{N}\sum fx^2 = \frac{1}{2^n}\left[1^2.{}^nC_1 + 2^2.{}^nC_2 + 3^2.{}^nC_2 + \ldots + n^2.{}^nC_n\right]$$

$$= \frac{n}{2^n}\left[1 + 2(n-1) + \frac{3}{2}(n-1)(n-2) + \ldots + n\right]$$

$$= \frac{n}{2^n}\left(\left\{1 + (n-1) + \frac{(n-1)(n-2)}{2!} + \ldots + 1\right\}\right.$$
$$\left. + \{(n-1) + (n-1)(n-2) + \ldots + (n-1)\}\right)$$

$$= \frac{n}{2^n}\left[\left({}^{n-1}C_0 + {}^{n-1}C_1 + {}^{n-1}C_2 + \ldots + {}^{n-1}C_{n-1}\right)\right.$$
$$\left. + \{(n-1)({}^{n-2}C_0 + {}^{n-2}C_1 + \ldots + {}^{n-2}C_{n-2})\}\right]$$

$$= \frac{n}{2^n}\left[(1+1)^{n-1} + (n-1)(1+1)^{n-2}\right] = \frac{n(n+1)}{4}$$

$$\therefore \qquad \sigma^2 = \frac{n(n+1)}{4} - \frac{n^2}{4} = \frac{n}{4}.$$

25. (*a*) Let r be the range and s be the standard deviation of a set of observations $x_1, x_2, \ldots, x_n$; then prove by general reasoning or otherwise that $s \leq r$.

Hint. Since $x_i - \bar{x} \leq r$, $i = 1, 2, \ldots, n$, we have

$$s^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2 \leq \frac{1}{N}\sum_{i=1}^{n} f_i (r^2)$$

$$\Rightarrow \qquad s^2 \leq r^2 \frac{1}{N}\sum_{i=1}^{n} f_i = r^2 \quad \Rightarrow \quad s \leq r$$

(*b*) Let r be the range and

$$S = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}}$$

be the standard deviation of a set of observations $x_1, x_2, \ldots, x_n$, then prove that

$$S \leq r\left(\frac{n}{n-1}\right)^{\frac{1}{2}}$$ **[Punjab Univ. B.Sc (Stat. Hons.), 1993]**

3·9. Moments. The rth moment of a variable x about any point $x = A$, usually denoted by μ_r' is given by

$$\mu_r' = \frac{1}{N}\sum_i f_i (x_i - A)^r, \quad \sum_i f_i = N \qquad \ldots(3\cdot14)$$

$$= \frac{1}{N}\sum_i f_i d_i^r, \qquad \ldots(3\cdot14a)$$

where $d_i = x_i - A$.

The rth moment of a variable about the mean $\bar{x}$, usually denoted by μ_r is given by

$$\mu_r = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^r = \frac{1}{N}\sum_i f_i z_i^r \qquad \ldots(3\cdot15)$$

where $\quad z_i = x_i - \bar{x}$.

In particular

$$\mu_0 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^0 = \frac{1}{N} \sum_i f_i = 1$$

and $\mu_1 = \frac{1}{N} \sum_i f_i (x_i - \bar{x}) = 0$, being the algebraic sum of deviations from the mean. Also

$$\mu_2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \sigma^2 \qquad \text{...(3·16)}$$

These results, viz., $\mu_0 = 1$, $\mu_1 = 0$, and $\mu_2 = \sigma^2$, are of fundamental importance and should be committed to memory.

We know that if $d_i = x_i - A$, then

$$\bar{x} = A + \frac{1}{N} \sum_i f_i d_i = A + \mu_1' \qquad \text{...(3·17)}$$

3·9·1. Relation between moments about mean in terms of moments about any point and *vice versa*.

We have

$$\mu_r = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^r = \frac{1}{N} \sum_i f_i (x_i - A + A - \bar{x})^r$$

$$= \frac{1}{N} \sum_i f_i (d_i + A - \bar{x})^r, \text{ where } d_i = x_i - A$$

Using (3·17), we get

$$\mu_r = \frac{1}{N} \sum_i f_i (d_i - \mu_1')^r$$

$$= \frac{1}{N} \sum f_i \left[d_i^r - {}^rC_1 d_i^{r-1} \mu_1' + {}^rC_2 d_i^{r-2} \mu_1'^2 - {}^rC_3 d_i^{r-3} \mu_1'^3 + \dots + (-1)^r \mu_1'^r \right]$$

...(3·18)

$$= \mu_r' - {}^rC_1 \mu_{r-1}' \mu_1' + {}^rC_2 \mu_{r-2}' \mu_1'^2 - \dots + (-1)^r \mu_1'^r \qquad \text{[On using (3·14a)]}$$

In particular, on putting r = 2, 3 and 4 in (3·18), we get

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 2\mu_1'^3 \qquad \text{...(3.19)}$$

$$\mu_4 = \mu_4' - 4\mu_3' \mu_1' + 6\mu_2' \mu_1'^2 - 3\mu_1'^4$$

Conversely,

$$\mu_r' = \frac{1}{N} \sum_i f_i (x_i - A)^r = \frac{1}{N} \sum_i f_i (x_i - \bar{x} + \bar{x} - A)^r$$

$$= \frac{1}{N} \sum_i f_i (z_i + \mu_1')^r$$

where $x_i - \bar{x} = z_i$ and $\bar{x} = A + \mu_1'$

Thus $\quad \mu_r' = \dfrac{1}{N} \sum_i f_i \left(z_i^r + {}^rC_1 z_i^{r-1} \mu_1' + {}^rC_2 z_i^{r-2} \mu_1'^2 + \dots + \mu_1'^r \right)$

$$= \mu_r + {}^rC_1 \mu_{r-1} \mu_1' + {}^rC_2 \mu_{r-2} \mu_1'^2 + \dots + \mu_1'^r. \text{ [From (3·15)]}$$

In particular, putting $r = 2$, 3 and 4 and noting that $\mu_1 = 0$, we get

$$\mu_2' = \mu_2 + \mu_1'^2$$
$$\mu_3' = \mu_3 + 3\mu_2 \mu_1' + \mu_1'^3 \qquad \dots (3·20)$$
$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2 \mu_1'^2 + \mu_1'^4$$

These formulae enable us to find the moments about any point, once the mean and the moments about mean are known.

3.9.2 Effect of Change of Origin and Scale on Moments.

Let $u = \dfrac{x - A}{h}$, so that $x = A + hu$, $\bar{x} = A + h\bar{u}$ and $x - \bar{x} = h(u - \bar{u})$

Thus, rth moment of x about any point $x = A$ is given by

$$\mu_r' = \frac{1}{N} \sum_i f_i (x_i - A)^r = \frac{1}{N} \sum_i f_i \ (hu_i)^r = h^r \cdot \frac{1}{N} \cdot \sum_i f_i u_i^r$$

And the rth moment of x about mean is

$$\mu_r = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^r = \frac{1}{N} \sum_i f_i [h(u_i - \bar{u})]^r$$
$$= h^r \frac{1}{N} \sum_i f_i (u_i - \bar{u})^r$$

Thus the rth moment of the variable x about mean is h^r times the rth moment of the variable u about its mean.

3·9·3. Sheppard's Corrections for Moments. In case of grouped frequency distribution, while calculating moments we assume that the frequencies are concentrated at the middle point of the class intervals. If the distribution is symmetrical or slightly symmetrical and the class intervals are not greater than one-twentieth of the range, this assumption is very nearly true. But since the assumption is not in general true, some error, called the 'grouping error', creeps into the calculation of the moments. W.F. Sheppard proved that if

(*i*) the frequency distribution is continuous, and

(*ii*) the frequency tapers off to zero in both directions,

the effect due to grouping at the mid-point of the intervals can be corrected by the following formulae, known as Sheppard's corrections :

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12} \qquad \dots (3·21)$$
$$\mu_3 \text{ (corrected)} = \mu_3$$
$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2} h^2 \mu_2 + \frac{7}{240} h^4$$

where h is the width of the class interval.

3·9·4. Charlier's Checks. The following identities

$\sum f(x+1) = \sum fx + N;\ \sum f(x+1)^2 = \sum fx^2 + 2\sum fx + N$

$\sum f(x+1)^3 = \sum fx^3 + 3\sum fx^2 + 3\sum fx + N$

$\sum f(x+1)^4 = \sum fx^4 + 4\sum fx^3 + 6\sum fx^2 + 4\sum fx + N,$

are often used in checking the accuracy in the calculation of first four moments and are known as Charlier's Checks.

3·10. Pearson's β and γ Coefficients. Karl Pearson defined the following four coefficients, based upon the first four moments about mean :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}\ ,\ \gamma_1 = +\sqrt{\beta_1} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2},\ \gamma_2 = \beta_2 - 3 \qquad \ldots (3\cdot22)$$

It may be pointed out that these coefficients are pure numbers independent of units of measurement. The practical utility of these coefficients is discussed in § 3·13 and § 3·14.

Remark. Sometimes, another coefficient based on moments, *viz*, Alpha (α) coefficient is used. Alpha coefficients are defined as :

$$\alpha_1 = \frac{\mu_1}{\sigma} = 0\ ,\ \alpha_2 = \frac{\mu_2}{\sigma^2} = 1,\ \alpha_3 = \frac{\mu^3}{\sigma^3} = \sqrt{\beta_1} = \gamma_1;\ \alpha_4 = \frac{\mu_4}{\sigma^4} = \beta_2$$

3·11. Factorial Moments. Factorial moment of order r about the origin of the frequency distribution $x_i \mid f_i$, $(i = 1, 2, \ldots n)$, is defined as

$$\mu_{(r)}' = \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i^{(r)} \qquad \ldots (3\cdot23)$$

where $x^{(r)} = x\,(x-1)\,(x-2)\ldots(x-r+1)$ and $N = \sum_{i=1}^{n} f_i$

Thus the factorial moment of order r about any point $x = a$ is given by

$$\mu_{(r)}' = \frac{1}{N}\sum_i f_i\ (x_i - a)^{(r)} \qquad \ldots (3\cdot24)$$

where $(x-a)^{(r)} = (x-a)\,(x-a-1)\ldots(x-a-r+1)$

In particular from (3·23), we have

$$\mu_{(1)}' = \frac{1}{N}\sum_i f_i\, x_i = \mu_1' \text{ (about origin)} = \text{Mean } (\bar{x})$$

$$\mu_{(2)}' = \frac{1}{N}\sum_i f_i\, x_i^{(2)} = \frac{1}{N}\sum_i f_i x_i\,(x_i - 1)$$

$$= \frac{1}{N}\sum_i f_i x_i^2 - \frac{1}{N}\sum_i f_i\, x_i = \mu_2' - \mu_1'$$

$$\mu_{(3)}' = \frac{1}{N}\sum_i f_i x_i^{(3)} = \frac{1}{N}\sum_i f_i x_i (x_i - 1)(x_i - 2)$$

$$= \frac{1}{N}\sum_i f_i x_i^3 - 3\,\frac{1}{N}\sum_i f_i x_i^2 + 2\,\frac{1}{N}\cdot\sum_i f_i x_i$$

$$= \mu_3' - 3\mu_2' + 2\mu_1'$$

$$\mu_{(4)}' = \frac{1}{N}\sum_i f_i x_i^{(4)} = \frac{1}{N}\sum_i f_i x_i (x_i - 1)(x_i - 2)(x_i - 3)$$

$$= \frac{1}{N}\sum_i f_i x_i (x_i^3 - 6x_i^2 + 11x_i - 6)$$

$$= \frac{1}{N}\sum_i f_i x_i^4 - 6.\,\frac{1}{N}\sum f_i x_i^3 + 11.\,\frac{1}{N}\sum f_i x_i^2 - 6.\,\frac{1}{N}\sum f_i x_i$$

$$= \mu_4' - 6\mu_3' + 11\mu_2' - 6\mu_1'$$

Conversely, we will get

$$\mu_1' = \mu_{(1)}'$$

$$\mu_2' = \mu_{(2)}' + \mu_{(1)}'$$

$$\mu_3' = \mu_{(3)}' + 3\mu_{(2)}' + \mu_{(1)}' \qquad \dots (3\cdot25)$$

$$\mu_4' = \mu_{(4)}' + 6\mu_{(3)}' + 7\mu_{(2)}' + \mu_{(1)}'$$

3·12. Absolute Moments. For the frequency distribution x_i / f_i $i = 1, 2, \dots n$, the rth absolute moment of the variable about the origin is given by

$$\frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i^r \right|, N = \sum f_i \qquad \dots (3\cdot 26)$$

where $\left| x_i^r \right|$ represents the absolute or modulus value of x_i^r.

The rth absolute moment of the variable about the mean $\bar{x}$ is given by

$$\frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i - \bar{x} \right|^r \qquad \dots (3\cdot26a)$$

Example 3·8. *The first four moments of a distribution about the value 4 of the variable are* $-1\cdot5$, 17, -30 *and* 108. *Find the moments about mean,* β_1 *and* β_2.

Find also the moments about (i) the origin, and (ii) the point $x = 2$.

Solution. In the usual notations, we are given $A = 4$ and

$$\mu_1' = -1\cdot5,\ \mu_2' = 17,\ \mu_3' = -30 \text{ and } \mu_4' = 108.$$

Moments about mean : $\mu_1 = 0$

$$\mu_2 = \mu_2' - \mu_1'^2 = 17 - (-1\cdot5)^2 = 17 - 2\cdot25 = 14\cdot75$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= -30 - 3 \times (17) \times (-1\cdot5) + 2(-1\cdot5)^3$$

$$= -30 + 76\cdot5 - 6\cdot75 = 39\cdot75$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 108 - 4(-30)(-1{\cdot}5) + 6(17)(-1{\cdot}5)^2 - 3(-1{\cdot}5)^4$$
$$= 108 - 180 + 229{\cdot}5 - 15{\cdot}1875 = 142{\cdot}3125$$

Hence $\qquad \beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{(39{\cdot}75)^2}{(14{\cdot}75)^3} = 0{\cdot}4924$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{(142{\cdot}3125)}{(14{\cdot}75)^2} = 0{\cdot}6541$$

Also $\qquad \bar{x} = A + \mu_1' = 4 + (-1{\cdot}5) = 2{\cdot}5$

Moments about origin. We have

$$\bar{x} = 2{\cdot}5,\ \mu_2 = 14{\cdot}75,\ \mu_3 = 39{\cdot}75 \text{ and } \mu_1 = 142{\cdot}31 \text{ (approx).}$$

We know $\bar{x} = A + \mu_1'$, where μ_1' is the first moment about the point $x = A$. Taking $A = 0$, we get the first moment about origin as $\mu_1' =$ mean $= 2{\cdot}5$.

Using (3·20), we get

$$\mu_2' = \mu_2 + \mu_1'^2 = 14{\cdot}75 + (2{\cdot}5)^2 = 14{\cdot}75 + 6{\cdot}25 = 21$$
$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3 = 39{\cdot}75 + 3(14{\cdot}75)(2{\cdot}5) + (2{\cdot}5)^3$$
$$= 39{\cdot}75 + 110{\cdot}625 + 15{\cdot}625 = 166$$
$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4$$
$$= 142{\cdot}3125 + 4(39{\cdot}75)(2{\cdot}5) + 6(14{\cdot}75)(2{\cdot}5)^2 + (2{\cdot}5)^4$$
$$= 142{\cdot}3125 + 397{\cdot}5 + 553{\cdot}125 + 39{\cdot}0625$$
$$= 1132.$$

Moments about the point $x = 2$. We have $\bar{x} = A + \mu_1'$. Taking $A = 2$, the first moment about the point $x = 2$ is

$$\mu_1' = \bar{x} - 2 = 2{\cdot}5 - 2 = 0{\cdot}5$$

Hence

$$\mu_2' = \mu_2 + \mu_1'^2 = 14{\cdot}75 + 0{\cdot}25 = 15$$
$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3 = 39{\cdot}75 + 3(14{\cdot}75)(0{\cdot}5) + (0{\cdot}5)^3$$
$$= 39{\cdot}75 + 22{\cdot}125 + 0{\cdot}125 = 62$$
$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4$$
$$= 142{\cdot}3125 + 4(39{\cdot}75)(0{\cdot}5) + 6(14{\cdot}75)(0{\cdot}5)^2 + (0{\cdot}5)^4$$
$$= 142{\cdot}3125 + 79{\cdot}5 + 22{\cdot}125 + 0{\cdot}0625$$
$$= 244$$

Example 3·9. *Calculate the first four moments of the following distribution about the mean and hence find* β_1 and β_2.

x:	0	1	2	3	4	5	6	7	8
f:	1	8	28	56	70	56	28	8	1

Solution. CALCULATION OF MOMENTS

x	f	$d = x - 4$	fd	fd^2	fd^3	fd^4
0	1	–4	–4	16	–64	256
1	8	–3	–24	72	–216	648
2	28	–2	–56	112	–224	448
3	56	–1	–56	56	–56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
Total	256	0	0	512	0	2,816

Moments about the points $x = 4$ are

$$\mu_1' = \frac{1}{N}\Sigma fd = 0,\ \mu_2' = \frac{1}{N}\Sigma fd^2 = \frac{512}{256} = 2,$$

$$\mu_3' = \frac{1}{N}\Sigma fd^3 = 0 \text{ and } \mu_4' = \frac{1}{N}\Sigma fd^4 = \frac{2816}{256} = 11$$

Moments about mean are :

$$\mu_1 = 0,\ \mu_2 = \mu_2' - \mu_1'^2 = 2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 11$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0,\ \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2{\cdot}75$$

Example 3·10 *For a distribution the mean is 10, variance is 16, γ_2 is + 1 and β_2 is 4. Obtain the first four moments about the orgin, i.e., zero. Comment upon the nature of distribution.*

Solution. We are given

Mean = 10, $\mu_2 = 16$, $\gamma_1 = +1$, $\beta_1 = 4$

First four moments about origin (μ_1', μ_2', μ_3', μ_4')

μ_1' = First moment about origin = Mean = 10

$$\mu_2 = \mu_2' - \mu_1'^2 \Rightarrow \mu_2' = \mu_2 + \mu_1'^2 \Rightarrow \mu_2' = 16 + 10^2 = 116$$

we have $\gamma_1 = +1 \Rightarrow \dfrac{\mu_3}{\mu_2^{3/2}} = 1$

$$\Rightarrow \mu_3 = \mu_2^{3/2} = (16)^{3/2} = 4^3 = 64$$

$$\therefore\ \mu_3 = \mu_3' - 3\mu_2'\mu_1' +-2\mu_1'^3$$

$$\Rightarrow \mu_3' = \mu_3 + 3\mu_2'\mu_1' - 2\mu_1'^3$$

$$= 64 + 3 \times 116 \times 10 - 2 \times 1000 = 3544 - 2000 = 1544$$

Now $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 4 \Rightarrow \mu_4 = 4 \times 16^2 = 1024$

and $\mu_4 = \mu'_4 - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - \mu_1'^4$

$\Rightarrow \quad \mu_4' = 1024 + 4 \times 1544 \times 10 - 6 \times 116 \times 100 + 3 \times 10000$
$= 92784 - 69600 = 23184.$

Comments of Nature of distribution : [c.f. § 3·13 and § 3·14] Since $\gamma_1 = +1$, the distribution is moderately positively skewed, i.e, if we draw the curve for the given distribution, it will have longer tail towards the right. Further since $\beta_2 = 4 > 3$, the distribution is leptokurtic, i.e., it will be more peaked than the normal curve.

Example 3·11. *If for a random variable x, the absolute moment of order k exists for ordinary k = 1, 2, ..., n–1, then the following inequalities*

(i). $\beta_k^{2k} \leq \beta_{k-1}^k \cdot \beta_{k+1}^k$, *(ii)* $\beta_k^{1/k} \leq \beta_{k+1}^{1/k+1}$

holds for k=1, 2, ..., n–1, where β_k *is the kth absolute moment about the origin.*

[Delhi Univ. B.Sc. (Stat.Hons.) 1989]

Solution. If $x_i \mid f_i$, $i = 1, 2, ..., n$ is the given frequency distribution, then

$$\beta_k = \frac{1}{N} \Sigma f_i \left| x_i^k \right| \qquad ...(1)$$

Let u and v be arbitrary real numbers, then the expression

$$\sum_{i=1}^{n} f_i \left[u \mid x_i^{(k-1)/2} \mid + v \mid x_i^{(k+1)/2} \mid \right]^2, \text{ is non-negative.}$$

$$\Rightarrow \sum_{i=1}^{n} f_i \left[u \mid x_i \mid^{(k-1)/2} + v \mid x_i \mid^{(k+1)/2} \right]^2 \geq 0$$

$$\Rightarrow \mu^2 \Sigma f_i |x_i|^{k-1} + v^2 \Sigma f_i |x_i|^{k+1} + 2uv \Sigma f_i |x_i|^k \geq 0$$

Dividing throughout by N and using relation (1), we get

$$u^2 \beta_{k-1} + v^2 \beta_{k+1} + 2uv\beta_k \geq 0, \; i.e., \; u^2 \beta_{k-1} + 2uv\beta_k + v^2 \beta_{k+1} \geq 0 \qquad ...(2)$$

We know that the condition for the expression $a x^2 + 2hxy + b y^2$ to be non - negative for all values of x and y is that

$$\begin{vmatrix} a & h \\ h & b \end{vmatrix} \geq 0$$

Using this result, we get from (2)

$$\begin{vmatrix} \beta_{k-1} & \beta_k \\ \beta_k & \beta_{k+1} \end{vmatrix} \geq 0$$

$$\Rightarrow \qquad \beta_{k-1} \cdot \beta_{k+1} - \beta_k^2 \geq 0 \qquad ...(3)$$

Raising both sides of (3) to power k, we get

$$\beta_k^{2k} \geq \beta_{k-1}^k \cdot \beta_{k+1}{}^k \qquad ...(4)$$

Putting k=1, 2, ..., k-1, k successively in (4), we get

$$\beta_1{}^2 \leq \beta_0 \beta_2$$
$$\beta_2^4 \leq \beta_1^2 \beta_3^2$$
$$\beta_3^6 \leq \beta_2^3 \beta_4^3$$
$$\cdots\cdots$$
$$\cdots\cdots$$
$$\beta_{k-1}^{2(k-1)} \leq \beta_{k-2}^{k-1} \cdot \beta_k^{k-1}$$

$$\beta_k^{2k} \le \beta_{k-1}^k \beta_{k+1}^k$$

Multiplying these inequalities and noting that $\beta_0 = 1$, we get

$$\beta_k^{k+1} \le \beta_{k+1}^k \text{ for } k = 1, 2, \ldots, n-1.$$

Raising both sides of the inequality to the power $\dfrac{1}{k(k+1)}$, we get

$$\beta_k^{1/k} \le \beta_{k+1}^{1/(k+1)} \qquad \ldots (5)$$

Remark. Result (5) shows that $\beta_k^{1/k}$ is an increasing function of k.

EXERCISE 3 (b)

1. (*a*) Define the raw and central moments of a frequency distribution. Obtain the relation between the central moments of order r in terms of the raw moments. What are Sheppard's corrections to the central moments ?

(*b*) Define moments. Establish the relationship between the moments about mean, *i.e.*, Central Moments in terms of moments about any arbitrary point and *vice versa.*

The first three moments of a distribution about the value 2 of the variable are 1, 16 and − 40. Show that the mean is 3, the variance is 15 and $\mu_3 = -86$. Also show that the first three moments about $x = 0$ are 3, 24 and 76.

(*c*) For a distribution the mean is 10, variance is 16, γ_1 is + 1 and β_2 is 4. Find the first four moments about the origin.

Ans. $\mu_1' = 10$, $\mu_2' = 116$, $\mu_3' = 1544$ and $\mu_4' = 23184$.

(*d*) (*i*) Define 'moment'. What is its use ? Express first four central moments in terms of moments about the origin. What is the effect of change of origin and scale on μ_3 ?

(*ii*) The first three moments of a distribution about the point $X = 7$ are 3, 11 and 15 respectively. Obtain mean, variance and β_1.

2. The first four moments of distribution about the value 5 of the variable are 2, 20, 40 and 50. Obtain as far as possible, the various characteristics of the distribution on the basis of the information given.

Ans. Mean = 7, $\mu_2 = 16$, $\mu_3 = -64$, $\mu_4 = 162$, $\beta_1 = 1$ and $\beta_2 = 0·63$.

3. (*a*) If the first four moments of a distribution about the value 5 are equal to − 4, 22, − 117 and 560, determine the corresponding moments (*i*) about the mean, (*ii*) about zero.

(*b*) What is Sheppard's correction? What will be the corrections for the first four moments ?

The first four moments of a distribution about $x = 4$ are 1, 4, 10, 45. Show that the mean is 5 and the variance is 3 and μ_3 and μ_4 are 0 and 26 respectively,

(*c*) In certain distribution, the first four moments about the point 4 are −1·5, 17, − 13 and 308. Calculate β_1 and β_2.

(*d*) The first four moments of a frequency distribution about the point 5 are −0.55, 4·46, − 0·43 and 68·52. Find β_1 and β_2.

Ans. $\mu_2 = 4·1575$, $\mu_3 = 6·5962$, $\mu_4 = 75·3944$, $\beta_1 = 0·6055$, $\beta_2 = 4·3619$.

4. (*a*) For the following data, calculate (*i*) Mean, (*ii*) Median, (*iii*) Semi-inter-quartile range, (*iv*) Coefficient of variation, and (*v*) β_1 and β_2 coefficients.

Wages in Rupees :	170—180	180—190	190—200	200—210	210—220	220—230	230—240	240—250
No. of Persons :	52	68	85	92	100	95	70	28

Ans. Mean = 209 (approx.); Median = 209·8; Q.D. = 15·8; σ = 19·7; *C.V.* = 9·4; $\beta_1 = 0{\cdot}003$; $\beta_2 = 26{\cdot}105$.

(*b*) Find the second, third and fourth central moments of the frequency distribution given below. Hence find (*i*) a measure of skewness (γ_1) and (*ii*) measure of kurtosis (γ_2).

Class Limits	*Frequency*
100·0 – 114·9	5
115·0 – 119·9	15
120·0 – 124·9	20
125·0 – 129·9	35
130·0 – 134·9	10
135·0 – 139·9	10
140·0 – 144·9	5

Also apply Sheppard's corrections for moments.

Ans. $\mu_2 = 2{\cdot}16$, $\mu_3 = 0{\cdot}804$, $\mu_4 = 12{\cdot}5232$

$$\gamma_1 = \sqrt{\beta_1} = 0{\cdot}25298;\ \gamma_2 = \beta_2 - 3 = -0{\cdot}317.$$

(*c*) The standard deviation of a symmetrical distribution is 5. What must be the value of the fourth moment about the mean in order that the distribution be (*i*) leptokurtic, (*ii*) mesokurtic, and (*iii*) platykurtic ?

Hint : $\mu_1 = \mu_3 = 0$ (because distribution is symmetrical).

$$\sigma = 5 \Rightarrow \sigma^2 = \mu_2 = 25$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{625}$$

(*i*) Distribution is leptokurtic if $\beta_2 > 3$, *i.e.*, if $\dfrac{\mu_4}{625} > 3 \Rightarrow \mu_4 > 1875$

(*ii*) Distribution is mesokurtic if $\beta_2 = 3 \Rightarrow$ if $\mu_4 = 1875$

(*iii*) Distribution is platykurtic if $\beta_2 < 3 \Rightarrow$ if $\mu_4 < 1875$

5. Show that for discrete distribution $\beta_2 > 1$.

[Allahabad Univ. M.A., 1993; Delhi Univ. B.Sc. (Stat. Hons), 1992]

Hint. We have to show that $\mu_4/\mu_2^2 > 1$, *i.e.*, $\mu_4 > \mu_2^2$. *If* x_i / f_i, $i = 1, 2, \ldots, n$, be the given discrete distribution, then we have to prove that

$$\frac{1}{N} \sum_i f_i\ (x_i - \bar{x})^4 > \left(\frac{1}{N} \sum_i f_i\ (x_i - \bar{x})^2\right)^2$$

Putting $(x_i - \bar{x})^2 = z_i$, we have to show that

$$\frac{1}{N} \sum_i f_i\, z_i^2 > \left(\frac{1}{N} \sum_i f_i\, z_i\right)^2$$

i.e., $\frac{1}{N} \sum_i f_i z_i^2 - \left(\frac{1}{N} \sum_i f_i z_i\right)^2 > 0$

i.e., $\sigma_z^2 > 0$,

which is always true, since variance is always positive.

Hence $\beta_2 > 1$.

6. (a) The scores in Economics of 250 candidates appearing at an examination have

Mean	$= \bar{x} = 39{\cdot}72$
Variance	$= \sigma^2 = 97{\cdot}80$
Third Central moment	$= \mu_3 = -114{\cdot}18$
Fourth central moment	$= \mu_4 = 28{,}396{\cdot}14$

It was later found on scrutiny that the score 61 of a candidate has been wrongly recorded as 51. Make necessary corrections in the given values of the mean and the central moment. **(Gujarat Univ. M.A., 1993)**

(b) For a distribution of 250 heights, calculations showed that the mean, standard deviation, β_1 and β_2 were 54 inches, 3 inches 0 and 3 inches respectively. It was, however, discovered on checking that the two items 64 and 50 in the original data were wrongly written in place of correct values 62 and 52 inches respectively. Calculate the correct frequency constants.

Ans. Correct Mean = 54, S.D. = 2·97, $\mu_3 = -2{\cdot}18$, $\mu_4 = 218{\cdot}42$, $\beta_1 = 0{\cdot}0070$ and $\beta_2 = 2{\cdot}81$

7. In calculating the moments of a frequency distribution based on 100 observations, the following results are obtained :

Mean = 9, Variance = 19, $\beta_1 = 0{\cdot}7$ (μ_3 + ive), $\beta_2 = 4$

But later on its was found that one observation 12 was read as 21. Obtain the correct value of the first four central moments.

Ans. Corrected mean = 8·91, $\mu_2 = 17{\cdot}64$, $\mu_3 = 57{\cdot}05$, $\mu_4 = 1257{\cdot}15$, $\beta_1 = 0{\cdot}59$ and $\beta_2 = 4{\cdot}04$.

8. (a) Show that if a range of six times the standard deviation covers at least 18 class intervals, Sheppard's correction will make a difference of less then 0·5 percent in the corrected value of the standard deviation.

Hint. If h is the magnitude of the class interval, then we want :

$$6\sigma > 18h \Rightarrow \sigma > 3h \Rightarrow h^2 < \frac{1}{9}\sigma^2 \Rightarrow -h^2 > -\frac{1}{9}\sigma^2$$

$$\therefore \quad \mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12} \geq \sigma^2 - \frac{1}{9 \times 12}\sigma^2 = \sigma^2\left(1 - \frac{1}{108}\right)$$

$$\Rightarrow \text{s.d. (corrected)} \geq \sigma\left(1 - \frac{1}{108}\right)^{1/2} \simeq \sigma\left(1 - \frac{1}{2} \times \frac{1}{108}\right)$$

$$\therefore \text{ Required adjustement} = \sigma - \sigma \text{ (corrected)} < \frac{\sigma}{216} < \frac{\sigma}{200} = \frac{1}{2}\ \%\ \textit{of s.d.}$$

(*b*) Show that, if the class intervals of a grouped distribution is less than one-third of the calculated standard deviation, Sheppard's adjustment makes a difference of less than $\frac{1}{2}$ % in the estimate of the standard deviation

9. (*a*) If ∂_r is the *r*th absolute moment about zero, use the mean value of

$$[u\,|\,x\,|^{(r-1)/2} + v\,|\,x\,|^{(r+1)/2}]^2$$

to show that

$$(\delta_r)^{2r} \le (\partial_{r-1})^r\,(\partial_{r+1})^r$$

From this derive the following inequalities :

(*i*) $(\partial_r)^{r+1} \le (\delta_{r+1})^r$, (*ii*) $(\partial_r)^{1/r} \le (\delta_{r+1})^{1/(r+1)}$

(*b*) For a random variable X moments of all order exist. Denoting by μ_j and ∂_j, the *j*th central moment and *j*th absolute moment respectively, show that

(*i*) $(\mu_{2j+1})^2 \le \mu_{2j}\;\mu_{2j+2}$,

(*ii*) $(\partial_j)^{1/j} \le (\partial_{j+1})^{1/(j+1)}$ **(Karnataka Univ. B.Sc., 1993)**

10. If β_1 and β_2 are the Pearsons's coefficients of skewness and Kurtosis respectively, show that $\beta_2 > \beta_1 + 1$. **(Bangalore Univ. B.Sc., 1993)**

3·13. Skewness. Literally, skewness means '*lack of symmetry*'. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. A distribution is said to be skewed if

(*i*) Mean, median and mode fall at different points,

i.e., Mean ≠ Median ≠ Mode,

(*ii*) Quartiles are not equidistant from median, and

(*iii*) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Measures of Skewness. Various measures of skewness are

(1) $S_k = M - M_d$ (2) $S_k = M - M_0$.

where M is the mean, M_d, the median and M_0, the mode of the distribution.

(3) $S_k = (Q_3 - M_d) - (M_d - Q_1)$.

These are the absolute measures of skewness. As in dispersion, for comparing two series we do not calculate these absolute measures but we calculate the relative measures called the *co-efficients of skewness* which are pure numbers independent of units of measurement. The following are the *coefficients of Skewness.*

1. *Prof. Karl Pearson's Coefficient of Skewness.*

$$S_k = \frac{(M - M_0)}{\sigma} \qquad \ldots (3\cdot27)$$

where σ is the standard deviation of the distribution.

If mode is ill-defined, then using the relation, $M_0 = 3M_d - 2M$, for a moderately asymmetrical distribution, we get

$$S_k = \frac{3(M - M_d)}{\sigma} \qquad \ldots (3\cdot27a)$$

The limits for Karl Pearson's coefficient of skewness are ± 3. In practice, these limits are rarely attained.

Skewness is positive if $M > M_0$ or $M > M_d$ and negative if $M < M_0$ or $M < M_d$.

II. *Prof. Bowley's Coefficient of Skewness.* Based on quartiles,

$$S_K = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} \qquad \dots (3\cdot28)$$

Remarks 1. Bowley's coefficient of skewness is also known as *Quartile coefficient of skewness* and is especially useful in situations where quartiles and median are used, *viz.*,

(*i*) When the mode is ill-defined and extreme observations are present in the data.

(*ii*) When the distribution has open end classes or unequal class intervals.

In these situations Pearson's coefficient of skewness cannot be used.

2. From (3·28), we observe that

$$S_k = 0\ ,\ if\ Q_3 - M_d = M_d - Q_1$$

This implies that for a symmetrical distribution ($S_k = 0$), median is equi-distant from the upper and lower quartiles. Moreover skewness is positive if :

$$Q_3 - M_d > M_d - Q_1 \quad \Rightarrow \quad Q_3 + Q_1 > 2M_d$$

and skewness is negative if

$$Q_3 - M_d < M_d - Q_1 \quad \Rightarrow \quad Q_3 + Q_1 < 2M_d$$

3. *Limits for Bowley's Coefficient of Skewness.* We know that for two real positive numbers a and b (*i.e.*, $a > 0$ and $b > 0$), the moduls value of the difference $(a - b)$ is always less than or equal to the modules value of the sum $(a + b)$, *i.e.*,

$$|a - b| \le |a + b| \quad \Rightarrow \quad \left|\frac{a - b}{a + b}\right| \le 1 \qquad \dots (*)$$

We also know that $(Q_3 - M_d)$ and $(M_d - Q_1)$ are both non-negative. Thus, taking $a = Q_3 - M_d$ and $b = M_d - Q_1$, in (*), we get

$$\left|\frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}\right| \le 1$$

$$\Rightarrow \qquad |S_k \text{ (Bowley)}| \le 1$$

$$\Rightarrow \qquad -1 \le Sk \text{ (Bowley)} \le 1.$$

Thus, Bowley's coefficient of skewness ranges from -1 to 1.

Further, we note from (3·28) that :

$$S_k = +1, \text{ if } M_d - Q_1 = 0, \textit{ i.e.}, \text{ if } M_d = Q_1$$

$$S_k = -1, \text{ if } Q_3 - M_d = 0, \textit{ i.e.}, \text{ if } Q_3 = M_d.$$

4. It should be clearly understood that the values of the coefficients of skewness obtained by Bowley's formula and Pearson's formula are not comparable, although in each case, $S_k = 0$, implies the absence of skewness, *i.e.*, the distribution is symmetrical. It may even happen that one of them gives positive skewness while the other gives negative skewness.

5. In Bowley's coefficient of skewness the disturbing factor of variation is eliminated by dividing the absolute measure of skewness, viz., $(Q_3 - Md) - (Md - Q_1)$ by the measure of dispersion $(Q_3 - Q_1)$, i.e., quartile range.

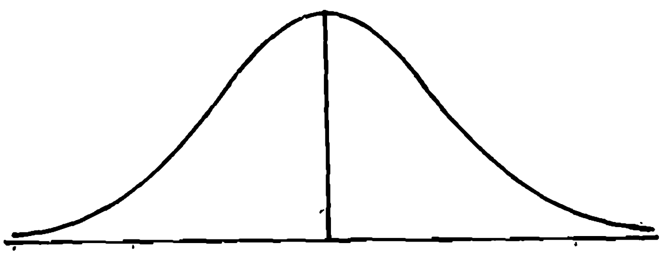
6. The only and perhaps quite serious limitations of this coefficient is that it is based only on the central 50% of the data and ignores the remaining 50% of the data towards the extremes.

III. Based upon moments, co-efficient of skewness is

$$S_k = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2\,(5\beta_2 - 6\beta_1 - 9)} \qquad \ldots(3{\cdot}29)$$

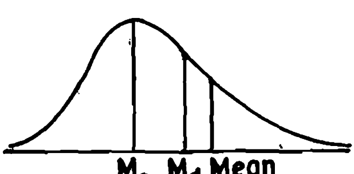
where symbols have their usual meaning. Thus $S_k = 0$ if either $\beta_1 = 0$ or $\beta_2 = -3$. But since $\beta_2 = \mu_4/\mu_2^2$, cannot be negative, $S_k = 0$ if and only if $\beta_1 = 0$. Thus for a symmetrical distribution $\beta_1 = 0$. In this respect β_1 is taken to be a measure of skewness. The co-efficient, in (3·29) is to be regarded as without sign.

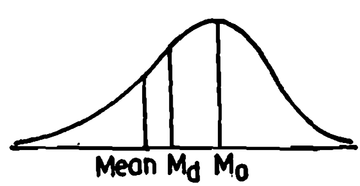
We observe in (3·27) and (3·28) that skewness can be positive as well as negative. The skewness is positive if the larger tail of the distribution lies towards



$\bar{x}$ (*Mean*) $= M_0 = M_d$
(Symmetrical Distribution)

the higher values of the variate (the right), i.e., if the curve drawn with the help of the given data is tretched more to the right than to the left and is negative



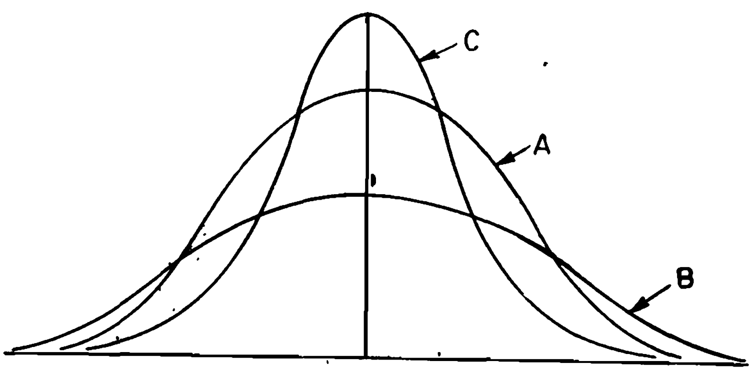


(Positively Skewed Distribution) (Negatively Skewed Distribution) in the contracy case.

3·14. Kurtosis. If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution as will be clear from the following figure in which all the three curves A, B and C are symmetrical about the mean 'm' and have the same range.

In addition to these measures we should know one more measure which Prof. Karl Pearson calls as the '*Convexity of curve*' or *Kurtosis*. Kurtosis enables us to have an idea about the flatness or peakedness of the curve. It is measured by the co-efficient β_2 or its derivation γ_2 given by

$$\beta_2 = \mu_4/\mu_2^2, \quad \gamma_2 = \beta_2 - 3$$



Curve of the type 'A' which is neither flat nor peaked is called the *normal curve or mesokurtic curve* and for such a curve $\beta_2 = 3$, i.e., $\gamma_2 = 0$. Curve of the type 'B' which is flatter than the normal curve is known as *platykurtic* and for such a curve $\beta_2 < 3$, *i.e.*, $\gamma_2 < 0$. Curve of the type 'C' which is more peaked than the normal curve is called *leptokurtic* and for such a curve $\beta_2 > 3$, i.e., $\gamma_2 > 0$.

EXERCISE 3 (c)

1. What do you understand by skewness ? How is it measured ? Distinguish clearly, by giving figures, between positive and negative skewness.

2. Explain the methods of measuring skewness and kurtosis of a frequency distribution.

3. Show that for any frequency distribution :

(i) Kurtosis is *greater* than unity.

(ii) Co-efficient of skewness is *less* than 1 numerically.

4. Why do we calculate in general, only the first four moments about mean of a distribution and not the higher moments ?

5. (a) Obtain Karl Pearsons's measure of skewness for the following data:

Values	Frequency	Values	Frequency
5 – 10	6	25 – 30	15
10 – 15	8	30 – 35	11
15 – 20	17	35 – 40	2
20 – 25	21		

(b) Assume that a firm has selected a random sample of 100 from its production line and has obtain the data shown in the table below :

Class interval	Frequency	Class interval	Frequency
130 – 134	3	150 – 154	19
135 – 139	12	155 – 159	12
140 – 144	21	160 – 164	5
145 – 149	28		

Compute the following :

(a) The arithmetic mean, (b) the standard deviation,

(c) Karl Pearson's coefficient of skewness.

Ans. (a) 147·2, (b) 7·2083 (c) 0·0711

6. (a) For the frequency distribution given below, calculate the coefficient of skewness based on quartiles.

Annual Sales (*Rs.* '000)	No. of Firms	Annual Sales (*Rs.* '000)	No. of firms
Less than 20	30	Less than 70	644
Less than 30	225	Less than 80	650
Less than 40	465	Less than 90	665
Less than 50	580	Less than 100	680
Less than 60	634		

(b) (i) Karl Pearsons's coefficient of skewness of a distribution is 0·32, its s.d. is 6·5 and mean is 29·6. Find the mode of the distribution.

(ii) If the mode of the above distribution is 24·8, what will be the s.d. ?

7. (a) In a frequency distribution, the co-efficient of skewness based upon the quartiles is 0·6. If the sum oif the upper and lower quartiles is 100 and median is 38, find the value of the upper and lower quartiles.

Hint. We are given

$$S_k = \frac{Q_3 + Q_1 - 2\ Md}{Q_3 - Q_1} = 0 \cdot 6 \qquad \ldots (*)$$

Also $Q_3 + Q_1 = 100$ and Median $= 38$

Subsituting (*), we get

$$\frac{100 - 2 \times 38}{Q_3 - Q_1} = 0 \cdot 6$$

$$\Rightarrow \qquad Q_3 - Q_1 = 40$$

Simplifying we get $\quad Q_1 = 30,\ Q_3 = 70$

(b) A frequency distribution gives the following results :

(i) C.V. = 5 (ii) Karl Pearsons's co-efficient of skewness = 0·5

(iii) $\sigma = 2$.

Find the mean and mode of the distribution.

(c) find the C.V. of a frequency distribution given that its mean is 120, mode is 123 and Karl Pearons's co-efficient of skewness is – 0·3.

Ans. C.V. = 8·33

(d) The first three moments of distribution about the value 2 are 1, 16 and 40 respectively. Examine the skewness of the distribution.

8. The first three moments about the origin 51 Kg. calculated from the data on the weights of 25 college students are

$$\mu_1' = +\ 0\cdot4 \text{ kg.}, \sqrt{\mu_2'} = 1\cdot2 \text{ kg. and } (\mu_3')^{1/2} = -\ 0\cdot25\text{kg.}$$

Determine the mean, the standard deviation and coefficient oif skewness.

9. The first three moments about the origin are given by

$$\mu_1' = \frac{n+1}{2},\ \mu_2' = \frac{(n+1)(2n+1)}{6} \text{ and } \mu_3' = \frac{n\ 9n+1)^2}{4}$$

Examine the skewness of the data.

10. Find out the kurtosis of the data given below :

Class interval	0 – 10	10 – 20	20 – 30	30 – 40
Frequency	1	3	4	2.

11. Data were obtained for distribution of passengers, entering Bombay local trains over time at intervals of 15 minutes for morning and evening rush hours separately, and the following results were obtained.

	Morning hours	*Evening hours*
Arithmetic mean (Peak Hours)	8 hrs. 38 min.	5 hrs. 40 min.
Standard deviation	38·5 min.	34·9 min.
Coefficient of skewness (in 15 min. unit)	- 0·32	+ 0·17
Kurtosis measure	2·0	2·2

Interpret the result and discuss giving reasons, whether you approve of the measure of 'peak hour'.

12. (a) The standard deviation of a symmertrical distribution is 5., What must be the value of the fourth moment about the mean in order that the distribution be (i) Leptokurtic, (ii) mesokurtic, and (iii) platykurtic.

Hint. $\mu_1 = \mu_3 = 0$ (Because distribution is symmetrical), $\sigma = 5 \Rightarrow \sigma^2 = \mu_2 = 25$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{625}$$

(i) Distt. is leptokurtic if $\beta_2 > 3$ i.e., if $\frac{\mu_4}{625} > 3 \Rightarrow \mu_4 > 1875$

(ii) Distt. is mesokurtic if $\beta_2 = 3 \Rightarrow$ if $\mu_4 < 1875$

(*iii*) Distt. is platykurtic if $\beta_2 < 3 \Rightarrow$ if $\mu_4 < 1875$.

(*b*) Find the second, third and fourth central moments of the frequency distribution given below., Hence, find (*i*) a measure of skewness, and (*ii*) a measure of kurtosis (γ_2).

Class limits	*Frequency*
110·0 — 114·9	5
115·0— 119·9	15
120·0— 124·9	20
125·0— 129·9	35
130·0— 134·9	10
135·0— 139·9	10
140·0— 144·9	5

Ans. $\mu_2 = 2{\cdot}16,\ \mu_3 = 0{\cdot}804,\ \mu_4 = 12{\cdot}5232.$

$$\gamma_1 = \sqrt{\beta_1} = 0{\cdot}25298\ ;\ \gamma_2 = \beta_2 - 3 = -\ 0{\cdot}317.$$

13. (*a*) Define Pearsonian coefficients β_1 and β_2 and discuss their utility in statistics. **[Delhi Univ. B.Sc. (Hons.), 1993]**

(*b*) What do you mean by skewness and kurtosis of a distribution ? Show that the Pearson's Beta coefficients satisfy the inequality $\beta_2 - \beta_1 - 1 \geq 0$. Also deduce that $\beta_2 \geq 1$. **(Delhi Univ. B.Sc. (Stat. Hons.), 1991)**

(*c*) Define the Pearson's coefficients γ_1 and γ_2 and discuss their utility in Statistics.

OBJECTIVE TYPE QUESTIONS

1. Match the correct parts to make a valid statement.

(*a*) Range	(*i*) $(Q_3 - Q_1)/2$
(*b*) Quartile Deviation	(*ii*) $\sqrt{\frac{1}{N}\ \Sigma\ f_i\ (x_i - \bar{x})^2}$
(*c*) Mean Deviation	(*iii*) $\frac{S.D.}{Mean} \times 100$
(*d*) Standard Deviation	(*iv*) $\frac{1}{N}\Sigma f_i \mid (x_i - \bar{x}) \mid$
(*e*) Coefficient of Variation	(*v*) $X_{max} - X_{min}$

II. Which value of '*a*' gives the minimum ?

(*i*) Mean square deviation from '*a*'

(*ii*) Mean deviation from '*a*'

III. Mean of 100 observations is 50 and S.D. is 10. What will be the new mean and S.D., if

(*i*) 5 is added to each observation,

(*ii*) each observation is multiplied by 3,

(*iii*) 5 is subtracted from each observation and then it is divided by 4?

IV. Fill in the blanks :

(*i*) (*a*) Absolute sum of deviation is minimum from...........

(b) Least value of root mean square deviation is ....

(ii) The sum of squares of deviations is least when measured from ....

(iii) The sum of 10 items is 12 and the sum of their squares is 16·9.

(iv) In any distribution, the standard deviation is always .... the mean deviation.

(v) The relationship between root mean square deviation and standard deviation σ is ....

(vi) If 25% of the items are less than 10 and 25% are more than 40, the coefficient of quartile deviation is ....

(vii) The median and standard deviation of a distribution are 20 and 4 respectively. If each item is increased by 2, the median will be .... and the new standard deviation will be ....

(viii) In a symmetric distribution, the mean and the mode are ....

(xi) In symmetric distribution, the upper and the lower quartiles are equi-distant from ....

(x) If the mean, mode and standard deviation of a frequency distribution are 41, 45 and 8 respectively, then its pearson's coefficient of skewness is ....

(xi) For a symmetrical distribution β_1 = ...

(xii) If $\beta_2 > 3$ the distribution is said to be ....

(xiii) For a symmetric distribution μ_2 = ...

μ_{2n+1} = ...

(xiv) If the mean and the mode of a given distribution are equal then its coefficient of skewness is ....

(xv) If the kurtosis of a distribution is 3, it is called .... distribution.

(xvi) In a perfectly symmetrical distribution 50% of items are above 60 and 75% items are below 75. Therefore, the coefficient of quartile deviation is .... and coefficient of skewness is ....

(xvii) Relation between β_1 and β_2 is given by ....

V. For the following questions given correct answers :

(i) Sum of absolute deviations about median is

(a) Least, (b) greatest, (c) zero, (d) equal.

(ii) The sum of squares of deviations is least when measured from

(a) Median, (b), (c) Mean, (d) Mode, (e) none of them.

(iii) In any discrete series (when all the values are not same) the relationship between M.D. about mean and S.D. is

(a) M.D. = S.D., (b) M.D. $\geq$ S.D., (c) M.D. < S.D.,

(d) M.D. $\leq$ S.D.

(e) None of these.

(iv) If each of a set of observations of a variable is multiplied by a constant (non-zero) value, the variance of the resultant variable.

(a) is unaltered, (b) increases (c) decreases, (d) is unknown.

(v) The appropriate measure whenever the extreme items are to be disregarded and when the distribution contains indefinite classes at the end is

(a) Median, (b) Mode, (c) Quartile deviation,
(d) Standard Deviation

(vi) A.M., G.M. and H.M. in any series are equal when

(a) the distribution is symmetric, (b) all the values are same,
(c) the distribution is positively skewed,
(d) the distribution is unimodal.

(vii) The limits for quartile coefficient of skewness are

(a) ± 3, (b) 0 and 3, (c) ± 1, (d) $\pm \infty$

(viii) 'The statement that the variance is equal to the second central moment'

(a) always true, (b) sometimes true, (c) never true,
(d) ambiguous.

(ix) The standard deviation of a distribution is 5. The value of the fourth central moment (μ_4), in order that the distribution be mesokurtic should be

(a) equal to 3, (b) greater than 1,875, (c) equal to 1,875,
(d) less than 1,875.

(x) In a frequency curve of scores the mode was found to be higher than the mean. This shows that the distribution is

(a) Symmetric, (b) negatively skewed, (c) positvely skewed,
(d) normal.

(xi) For any frequency distribution, the kurtosis is

(a) greater than 1, (b) less than 1, (c) equal to 1.

(xii) The measure of kurtosis is

(a) $\beta_2 = 0$, (b) $\beta_2 = 3$, (c) $\beta_2 = 4$.

(xiii) For the distribution

(a) $\mu_4 = 0$, (b) Median $= 0$,
(c) The distribution of x is symmetrical.

X:	−4	−3	−2	−1	0	1	2	3	4	Total
f:	2	4	5	7	10	7	5	4	2	46

(xiv) For a symmetric distribution

(a) $\mu_2 = 0$, (b) $\mu_2 > 0$; (c) $\mu_3 > 0$

VI. State which of the following statements are Ture and which False. In each of false statements given the correct statement.

(i) Mean, standard deviation and varaince have the same unit.

(ii) Standard deviation of every distribution is unique and always exists.

(iii) Median is the value of the variance which divides the total frequency it two equal parts.

(iv) Mean − Mode = 3 (mean − median) is often approximately satisified.

(v) Mean deviation $= \frac{4}{5}$ (standard deviation) is always satisfied.

(vi) $\beta_2 \geq 1$ is always satisfied

(vii) $\beta_1 = 0$ is a conclusive test for a distribution to be symmetrical.

Theory of Probability

4·1. Introduction. If an experiment is repeated under essentially homogeneous and similar conditions we generally come across two types of situations:

(i) The result or what is usually known as the *'outcome'* is unique or certain.

(ii) The result is not unique but may be one of the several possible outcomes.

The phenomena covered by (i) are known as *'deterministic'* or *'predictable'* phenomena. By a deterministic phenomenon we mean one in which the result can be predicted with certainty. For example :

(a) For a perfect gas,

$$V \propto \frac{1}{P} \quad i.e., \; PV = constant,$$

provided the temperature remains the same.

(b) The velocity 'v' of a particle after time 't' is given by

$$v = u + at$$

where u is the initial velocity and a is the acceleration. This equation uniquely determines v if the right-hand quantities are known.

(c) Ohm's Law, *viz.,* $C = \dfrac{E}{R}$

where C is the flow of current, E the potential difference between the two ends of the conductor and R the resistance, uniquely determines the value C as soon as E and R are given.

A deterministic model is defined as a model which stipulates that the conditions under which an experiment is performed determine the outcome of the experiment. For a number of situations the deterministic model suffices. However, there are phenomena [as covered by (ii) above] which do not lend themselves to deterministic approach and are known as *'unpredictable'* or *'probabilistic'* phenomena. For example :

(i) In tossing of a coin one is not sure if a head or tail will be obtained.

(ii) If a light tube has lasted for t hours, nothing can be said about its further life. It may fail to function any moment.

In such cases we talk of chance or probability which is taken to be a quantitative measure of certainty.

4·2. Short History. Galileo (1564-1642), an Italian mathematician, was the first to attempt at a quantitative measure of probability while dealing with some problems related to the theory of dice in gambling. But the first foundation of the mathematical theory if probability was laid in the mid-seventeenth century by two French mathematicians, B. Pascal (1623-1662) and P. Fermat (1601-1665), while solving a number of problems posed by French gambler and noble man Chevalier-De-Mere to Pascal. The famous *'problem of points'* posed by De-Mere to Pascal is : "Two persons play a game of chance. The person who first gains a certain number of points wins the stake. They stop playing before the game is completed. How is the stake to be decided on the basis of the number of points each has won?" The two mathematicians after a lengthy correspondence between themselves ultimately solved this problem and this correspondence laid the first foundation of the science of probability. Next stalwart in this field was J. Bernoulli (1654-1705) whose 'Treatise on Probability' was published posthumously by his nephew N. Bernoulli in 1713. De-Moivre (1667-1754) also did considerable work in this field and published his famous 'Doctrine of Chances' in 1718. Other main contributors are : T. Bayes (Inverse probability), P.S. Laplace (1749-1827) who after extensive research over a number of years finally published 'Theoric analytique des probabilities' in 1812. In addition to these, other outstanding contributors are Levy, Mises and R.A. Fisher.

Russian mathematicians also have made very valuable contributions to the modern theory of probability. Chief contributors, to mention only a few of them are : Chebyshev (1821-94) who founded the Russian School of Statisticians; A. Markoff (1856-1922); Liapounoff (Central Limit Theorem); A. Khintchine (Law of Large Numbers) and A. Kolmogorov, who axiomised the calculus of probability.

4·3. Definitions of Various Terms. In this section we will define and explain the various terms which are used in the definition of probability.

Trial and Event. Consider an experiment which, though repeated under essentially identical conditions, does not give unique results but may result in any one of the several possible outcomes. The experiment is known as a *trial* and the outcomes are known as *events* or *cases*. For example :

(i) Throwing of a die is a trial and getting 1(or 2 or 3, ...or 6) is an event.

(ii) Tossing of a coin is a trial and getting head (H) or tail (T) is an event.

(iii) Drawing two cards from a pack of well-shuffled cards is a trial and getting a king and a queen are events.

Exhaustive Events. The total number of possible outcomes in any trial is known as exhaustive events or exhaustive cases. For example :

(i) In tossing of a coin there are two exhaustive cases, *viz.*, head and tail (the possibility of the coin standing on an edge being ignored).

(ii) In throwing of a die, there are six exhaustive cases since any one of the 6 faces 1, 2, ...,6 may come uppermost.

(iii) In drawing two cards from a pack of cards the exhaustive number of cases is ${}^{52}C_2$, since 2 cards can be drawn out of 52 cards in ${}^{52}C_2$ ways.

(iv) In throwing of two dice, the exhaustive number of cases is $6^2 = 36$, since any of the 6 numbers 1 to 6 on the first die can be associated with any of the six numbers on the other die.

In general in throwing of n dice the exhaustive number of cases is 6^n.

Favourable Events or Cases. The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event. For example,

(i) In drawing a card from a pack of cards the number of cases favourable to drawing of an ace is 4, for drawing a spade is 13 and for drawing a red card is 26.

(ii) In throwing of two dice, the number of cases favourable to getting the sum 5 is : (1,4) (4,1) (2,3) (3,2), *i.e.*, 4.

Mutually exclusive events. Events are said to be *mutually exclusive* or *incompatible* if the happening of any one of them precludes the happening of all the others, *i.e.*, if no two or more of them can happen simultaneously in the same trial. For example :

(i) In throwing a die all the 6 faces numbered 1 to 6 are mutually exclusive since if any one of these faces comes, the possibility of others, in the same trial, is ruled out.

(ii) Similarly in tossing a coin the events head and tail are mutually exclusive.

Equally likely events. Outcomes of a trial are set to be equally likely if taking into consideration all the relevant evidences, there is no reason to expect one in preference to the others. For example

(i) In tossing an unbiased or uniform coin, head or tail are equally likely events.

(ii) In throwing an unbiased die, all the six faces are equally likely to come.

Independent events. Several events are said to be independent if the happening (or non-happening) of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events. For example

(i) In tossing an unbiased coin the event of getting a head in the first toss is independent of getting a head in the second, third and subsequent throws.

(ii) If we draw a card from a pack of well-shuffled cards and replace it before drawing the second card, the result of the second draw is independent of the first draw. But, however, if the first card drawn is not replaced then the second draw is dependent on the first draw.

4·3·1. Mathematical or Classical or 'a priori' Probabality

Definition. If a trial results in n exhaustive, mutually exclusive and equally likely cases and m of them are favourable to the happening of an event E, then the probability 'p' of happening of E is given by

$$p = P(E) = \frac{\textit{Favourable number of cases}}{\textit{Exhaustive number of cases}} = \frac{m}{n} \quad \text{....(4·1)}$$

Sometimes we express (4·1) by saying that 'the odds in favour of E are m : $(n-m)$ or the odds against E are $(n-m) : n$.'

Since the number of cases favourable to the 'non-happening' of the event E are $(n-m)$, the probability 'q' that E will not happen is given by

$$q=\frac{n-m}{n}=1-\frac{m}{n}=1-p \quad \Rightarrow \quad p+q=1 \qquad \ldots(4\cdot1a)$$

Obviously p as well as q are non-negative and cannot exceed unity, *i.e,* $0\leq p\leq 1$, $0\leq q\leq 1$.

Remarks. 1. Probability 'p' of the happening of an event is also known as the probability of success and the probability 'q' of the non- happening of the event as the probability of failure.

2. If $P(E)=1$, E is called a *certain event* and if $P(E)=0$, E is called an *impossible* event.

3. Limitations of Classical Definition. This definition of Classical Probability breaks down in the following cases :

(i) If the various outcomes of the trial are not equally likely or equally probable. For example, the probability that a candidate will pass in a certain test is not 50% since the two possible outcomes, *viz.,* sucess and failure (excluding the possibility of a compartment) are not equally likely.

(ii) If the exhaustive number of cases in a trial is infinite.

4·3·2. Statistical or Empirical Probability

Definition (Von Mises). *If a trial is repeated a number of times under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event happens to the number of trials, as the number of trials become indefinitely large, is called the probability of happening of the event. (It is assumed that the limit is finite and unique).*

Symbolically, if in n trials an event E happens m times, then the probability 'p' of the happening of E is given by

$$p=P(E)=\lim_{n\to\infty}\frac{m}{n} \qquad \ldots(4\cdot2)$$

Example 4·1. *What is the chance that a leap year selected at random will contain 53 Sundays?*

Solution. In a leap year (which consists of 366 days) there are 52 complete weeks and 2 days over. The following are the possible combinations for these two 'over' days:

(i) Sunday and Monday, *(ii)* Monday and Tuesday, *(iii)* Tuesday and Wednesday, *(iv)* Wednesday and Thursday, *(v)* Thursday and Friday, *(vi)* Friday and Saturday, and *(vii)* Saturday and Sunday.

In order that a leap year selected at random should contain 53 Sundays, one of the two 'over' days must be Sunday. Since out of the above 7 possibilities, 2 *viz.,* *(i)* and *(vii)*, are favourable to this event,

$$\therefore \qquad \text{Required probability}=\frac{2}{7}$$

Example 4·2. *A bag contains 3 red, 6 white and 7 blue balls. What is the probability that two balls drawn are white and blue?*

Solution. Total number of balls $= 3+6+7=16$.

Now, out of 16 balls, 2 can be drawn in ${}^{16}C_2$ ways.

$\therefore$ Exhaustive number of cases $= {}^{16}C_2 = \dfrac{16 \times 15}{2} = 120.$

Out of 6 white balls 1 ball can be drawn in ${}^{6}C_1$ ways and out of 7 blue balls 1 ball can be drawn in ${}^{7}C_1$ ways. Since each of the former cases can be associated with each of the latter cases, total number of favourable cases is : ${}^{6}C_1 \times {}^{7}C_1 = 6 \times 7 = 42$.

$$\therefore \quad \text{Required probability} = \frac{42}{120} = \frac{7}{20}.$$

Example 4·3. (a) *Two cards are drawn at random from a well-shuffled pack of 52 cards. Show that the chance of drawing two aces is 1/221.*

(b) From a pack of 52 cards, three are drawn at random. Find the chance that they are a king, a queen and a knave.

(c) Four cards are drawn from a pack of cards. Find the probability that

(i) all are diamond, (ii) there is one card of each suit, and (iii) there are two spades and two hearts.

Solution. *(a)* From a pack of 52 cards 2 cards can be drawn in ${}^{52}C_2$ ways, all being equally likely.

$\therefore$ Exhaustive number of cases $= {}^{52}C_2$

In a pack there are 4 aces and therefore 2 aces can be drawn in ${}^{4}C_2$ ways.

$$\therefore \quad \text{Required probability} = \frac{{}^{4}C_2}{{}^{52}C_2} = \frac{4 \times 3}{2} \times \frac{2}{52 \times 51} = \frac{1}{221}$$

(b) Exhaustive number of cases $= {}^{52}C_3$

A pack of cards contains 4 kings, 4 queens and 4 knaves. A king, a queen and a knave can each be drawn in ${}^{4}C_1$ ways and since each way of drawing a king can be associated with each of the ways of drawing a queen and a knave, the total number of favourable cases $= {}^{4}C_1 \times {}^{4}C_1 \times {}^{4}C_1$

$$\therefore \quad \text{Required probability} = \frac{{}^{4}C_1 \times {}^{4}C_1 \times {}^{4}C_1}{{}^{52}C_3} = \frac{4 \times 4 \times 4 \times 6}{52 \times 51 \times 50} = \frac{16}{5525}$$

(c) Exhaustive number of cases $= {}^{52}C_4$.

(i) Required probability $= \dfrac{{}^{13}C_4}{{}^{52}C_4}$

(ii) Required probability $= \dfrac{{}^{13}C_1 \times {}^{13}C_1 \times {}^{13}C_1 \times {}^{13}C_1}{{}^{52}C_4}$

(iii) Required probability $= \dfrac{{}^{13}C_2 \times {}^{13}C_2}{{}^{52}C_4}$

Example 4·4. *What is the probability of getting 9 cards of the same suit in one hand at a game of bridge?*

Solution. One hand in a game of bridge consists of 13 cards.

$\therefore$ Exhaustive number of cases $= {}^{52}C_{13}$

Number of ways in which, in one hand, a particular player gets 9 cards of one suit are ${}^{13}C_9$ and the number of ways in which the remaining 4 cards are of some other suit are ${}^{39}C_4$. Since there are 4 suits in a pack of cards, total number of favourable cases $= 4 \times {}^{13}C_9 \times {}^{39}C_4$.

$$\therefore \qquad \text{Required probability} = \frac{4 \times {}^{13}C_9 \times {}^{39}C_4}{{}^{52}C_{13}}$$

Example 4·5. *(a) Among the digits 1, 2, 3, 4, 5, at first one is chosen and then a second selection is made among the remaining four digits. Assuming that all twenty possible outcomes have equal probabilities, find the probability that an odd digit will be selected (i) the first time, (ii) the second time, and (iii) both times.*

(b) From 25 tickets, marked with the first 25 numerals, one is drawn at random. Find the chance that

(i) it is a multiple of 5 or 7,

(ii) it is a multiple of 3 or 7.

Solution. *(a)* Total number of cases $= 5 \times 4 = 20$

(i) Now there are 12 cases in which the first digit drawn is odd, *viz.*, (1, 2), (1, 3), (1, 4), (1, 5), (3, 1), (3, 2), (3, 4), (3, 5), (5, 1), (5, 2), (5, 3) and (5, 4).

$\therefore$ The probability that the first digit drawn is odd

$$= \frac{12}{20} = \frac{3}{5}$$

(ii) Also there are 12 cases in which the second digit drawn is odd, *viz.*, (2, 1), (3, 1), (4, 1), (5, 1), (1, 3), (2, 3), (4, 3), (5, 3), (1, 5), (2, 5), (3, 5) and (4, 5).

$\therefore$ The probability that the second digit drawn is odd

$$= \frac{12}{20} = \frac{3}{5}$$

(iii) There are six cases in which both the digits drawn are odd, *viz.*, (1, 3), (1, 5), (3, 1), (3, 5), (5, 1) and (5, 3).

$\therefore$ The probability that both the digits drawn are odd

$$= \frac{6}{20} = \frac{3}{10}$$

(b) (i) Numbers (out of the first 25 numerals) which are multiples of 5 are 5, 10, 15, 20 and 25, *i.e.*, 5 in all and the numbers which are multiples of 7 are 7, 14 and 21, *i.e.*, 3 in all. Hence required number of favourable cases are 5+3=8.

$$\therefore \quad \text{Required probability} = \frac{8}{25}$$

(ii) Numbers (among the first 25 numerals) which are multiples of 3 are 3, 6, 9, 12, 15, 18, 21, 24, *i.e.*, 8 in all, and the numbers which are multiples of 7 are 7,

14, 21, *i.e.*, 3 in all. Since the number 21 is common in both the cases, the required number of distinct favourable cases is $8 + 3 - 1 = 10$.

$$\therefore \qquad \text{Required probability} = \frac{10}{25} = \frac{2}{5}$$

Example 4·6. *A committee of 4 people is to be appointed from 3 officers of the production department, 4 officers of the purchase department, two officers of the sales department and 1 chartered accountant. Find the probability of forming the committee in the following manner:*

(i) There must be one from each category.

(ii) It should have at least one from the purchase department.

(iii) The chartered accountant must be in the committee.

Solution. There are 3+4+2+1=10 persons in all and a committee of 4 people can be formed out of them in ${}^{10}C_4$ ways. Hence exhaustive number of cases is

$${}^{10}C_4 = \frac{10 \times 9 \times 8 \times 7}{4\,!} = 210$$

(i) Favourable number of cases for the committee to consist of 4 members, one from each category is :

$${}^{4}C_1 \times {}^{3}C_1 \times {}^{2}C_1 \times 1 = 4 \times 3 \times 2 = 24$$

$$\therefore \qquad \text{Required probability} = \frac{24}{210} = \frac{8}{70}$$

(ii) P [Committee has at least one purchase officer]

$= 1 - P$ [Committee has no purchase officer]

In order that the committee has no purchase officer, all the 4 members are to be selected from amongst officers of production department, sales department and chartered accountant, *i.e.*, out of 3+2+1=6 members and this can be done in ${}^{6}C_4 = \dfrac{6 \times 5}{1 \times 2} = 15$ ways. Hence

$$P\,[\text{ Committee has no purchase officer }] = \frac{15}{210} = \frac{1}{14}$$

$$\therefore \quad P\,[\text{ Committee has at least one purchase officer }] = 1 - \frac{1}{14} = \frac{13}{14}$$

(iii) Favourable number of cases that the committee consists of a chartered accountant as a member and three others are :

$$1 \times {}^{9}C_3 = \frac{9 \times 8 \times 7}{1 \times 2 \times 3} = 84 \text{ ways,}$$

since a chartered accountant can be selected out of one chartered accountant in only 1 way and the remaining 3 members can be selected out of the remaining $10 - 1 = 9$ persons in ${}^{9}C_3$ ways. Hence the required probability $= \frac{84}{210} = \frac{2}{5}$.

Example 4·7. *(a) If the letters of the word 'REGULATIONS' be arranged at random, what is the chance that there will be exactly 4 letters between R and E?*

(b) What is the probability that four S's come consecutively in the word 'MISSISSIPPI'?

Solution. *(a)* The word '*REGULATIONS*' consists of 11 letters. The two letters R and E can occupy ${}^{11}P_2$, *i.e.*, $11 \times 10 = 110$ positions.

The number of ways in which there will be exactly 4 letters between R and E are enumerated below:

(i) R is in the 1st place and E is in the 6th place.

(ii) R is in the 2nd place and E is in the 7th place.

... ...

... ...

... ...

(vi) R is in the 6th place and E is in the 11th place.

Since R and E can interchange their positions, the required number of favourable cases is $2 \times 6 = 12$

$\therefore$ The required probability $= \dfrac{12}{110} = \dfrac{6}{55}$.

(b) Total number of permutations of the 11 letters of the word '*MISSISSIPPI*', in which 4 are of one kind *(viz., S)*, 4 of other kind *(viz., I)*, 2 of third kind *(viz., P)* and 1 of fourth kind *(viz., M)* are

$$\frac{11\,!}{4\,!\;4\,!\;2\,!\;1\,!}$$

Following are the 8 possible combinations of 4 S's coming consecutively:

(i)	S	S	S	S						
(ii)	—	S	S	S	S					
(iii)	—	—	S	S	S	S				
⋮			⋮			⋮				
(viii)	—	—	—	—	—	—	—	S	S	S	S

Since in each of the above cases, the total number of arrangements of the remaining 7 letters, *viz.*, *MIIIPPI* of which 4 are of one kind, 2 of other kind and one of third kind are $\dfrac{7\,!}{4\,!\;2\,!\;1\,!}$, the required number of favourable cases

$$= \frac{8 \times 7\,!}{4\,!\;2\,!\;1\,!}$$

$$\therefore \quad \text{Required probability} = \frac{8 \times 7\,!}{4\,!\;2\,!\;1\,!} \div \frac{11\,!}{4\,!\;4\,!\;2\,!\;1\,!}$$

$$= \frac{8 \times 7\,! \times 4\,!}{11\,!} = \frac{4}{165}$$

Example 4·8. *Each coefficient in the equation $ax^2 + bx + c = 0$ is determined by throwing an ordinary die. Find the probability that the equation will have real roots.* **[Madras Univ. B. Sc. (Stat. Main), 1992]**

Solution. The roots of the equation $ax^2 + bx + c = 0$ ...(*)
will be real if its discriminant is non-negative, *i.e.*, if

$$b^2 - 4ac \geq 0 \quad \Rightarrow \quad b^2 \geq 4ac$$

Since each co-efficient in equation (*) is determined by throwing an ordinary die, each of the co-efficients a, b and c can take the values from 1 to 6.

$\therefore$ Total number of possible outcomes (all being equally likely)

$$= 6 \times 6 \times 6 = 216$$

The number of favourable cases can be enumerated as follows:

ac		*a*	*c*	*4ac*	*b* (so that $b^2 \geq 4ac$)	*No. of cases*
1		1	1	4	2, 3, 4, 5	$1 \times 5 = 5$
2	(*i*)	1	2	8	3, 4, 5, 6	$2 \times 4 = 8$
	(*ii*)	2	1			
3	(*i*)	1	3	12	4, 5, 6	$2 \times 3 = 6$
	(*ii*)	3	1			
4	(*i*)	1	4	16	4, 5, 6	$3 \times 3 = 9$
	(*ii*)	4	1			
	(*iii*)	2	2			
5	(*i*)	1	5	20	5, 6	$2 \times 2 = 4$
	(*ii*)	5	1			
6	(*i*)	1	6	24	5, 6	$4 \times 2 = 8$
	(*ii*)	6	1			
	(*iii*)	3	2			
	(*iv*)	2	3			
7	(*ac* = 7 *is not possible*)					
8	(*i*)	2	4	32	6	$2 \times 1 = 2$
	(*ii*)	4	2			
9		3	3	36	6	1
						Total = 43

Since $b^2 \geq 4ac$ and since the maximum value of b^2 is 36, $ac = 10, 11, 12, \ldots$ etc. is not possible.

Hence total number of favourable cases = 43.

$\therefore$ Required probability $= \dfrac{43}{216}$

Example 4·9. *The sum of two non-negative quantities is equal to* $2n$. *Find the chance that their product is not less than* $\frac{3}{4}$ *times their greatest product.*

Solution. Let $x > 0$ and $y > 0$ be the given quantities so that $x + y = 2n$.

We know that the product of two positive quantities whose sum is constant is greatest when the quantities are equal. Thus the product of x and y is maximum when $x = y = n$.

$\therefore$ Maximum product $= n \cdot n = n^2$

Now $P\left[xy \not< \frac{3}{4}n^2\right] = P\left[xy \geq \frac{3}{4}n^2\right] = P\left[x(2n-x) \geq \frac{3}{4}n^2\right]$

$$= P\left[(4x^2 - 8nx + 3n^2) \leq 0\right]$$

$$= P\left[(2x-3n)(2x-n) \leq 0\right]$$

$$= P\left[x \text{ lies between } \frac{n}{2} \text{ and } \frac{3n}{2}\right]$$

$\therefore$ Favourable range $= \frac{3n}{2} - \frac{n}{2} = n$

Total range $= 2n$

$\therefore$ Required probability $= \frac{n}{2n} = \frac{1}{2}$

Example 4·10. *Out of (2n+1) tickets consecutively numbered three are drawn at random. Find the chance that the numbers on them are in A.P.*

[Calicut Univ. B.Sc., 1991; Delhi Univ. B.Sc.(Stat. Hons.), 1992]

Solution. Since out of $(2n+1)$ tickets, 3 tickets can be drawn in ${}^{2n+1}C_3$ ways,

$$\text{Exhaustive number of cases} = {}^{2n+1}C_3 = \frac{(2n+1)\,2n\,(2n-1)}{3\,!}$$

$$= \frac{n(4n^2-1)}{3}$$

To find the favourable number of cases we are to enumerate all the cases in which the numbers on the drawn tickets are in A.P with common difference, (say $d = 1, 2, 3, \ldots, n-1, n$).

If $d = 1$, the possible cases are as follows:

$$\left.\begin{array}{ccc} 1, & 2, & 3 \\ 2, & 3, & 4 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 2n-1, & n, & 2n+1 \end{array}\right\}, \text{ i.e., } (2n-1) \text{ cases in all}$$

If $d = 2$, the possible cases are as follows :

$$\left.\begin{array}{ccc} 1, & 3, & 5 \\ 2, & 4, & 6 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 2n-3, & 2n-1, & 2n+1 \end{array}\right\}, \text{ i.e., } (2n-3) \text{ cases in all}$$

and so on.

If $d = n-1$, the possible cases are as follows:

$$\left.\begin{array}{ll} 1, n, & 2n-1 \\ 2, n+1, & 2n \\ 3, n+2, & 2n+1 \end{array}\right\}, \text{ i.e., 3 cases in all}$$

If $d = n$, there is only one case, *viz.*, $(1, n+1, 2n+1)$.

Hence total number of favourable cases

$$= (2n-1) + (2n-3)) + ... + 5 + 3 + 1$$
$$= 1 + 3 + 5 + ... + (2n-1),$$

which is a series in A.P. with $d = 2$ and n terms.

$\therefore$ Number of favourable cases $= \frac{n}{2}[1 + (2n-1)] = n^2$

$\therefore$ Required probability $= \dfrac{n^2}{n\,(4n^2-1)/3} = \dfrac{3n}{(4n^2-1)}$

EXCERCISE 4 (a)

1. (*a*) Give the classical and statistical definitions of probability. What are the objections raised in these definitions?

[Delhi Univ. B.Sc. (Stat. Hons.), 1988, 1985]

(*b*) When are a number of cases said to be equally likely? Give an example each of the following :

(*i*) the equally likely cases,

(*ii*) four cases which are not equally likely, and

(*iii*) five cases in which one case is more likely than the other four.

(*c*) What is meant by mutually exclusive events? Give an example of

(*i*) three mutually exclusive events,

(*ii*) three events which are not mutually exclusive.

[Meerut Univ. B.Sc. (Stat.), 1987]

(*d*) Can

(*i*) events be mutually exclusive and exhaustive?

(*ii*) events be exhaustive and indepenent?

(*iii*) events be mutually exclusive and independent?

(*iv*) events be mutually exhaustive, exclusive and independent?

2. (*a*) Prove that the probability of obtaining a total of 9 in a single throw with two dice is one by nine.

(*b*) Prove that in a single throw with a pair of dice the probability of getting the sum of 7 is equal to 1/6 and the probability of getting the sum of 10 is equal to 1/12.

(*c*) Show that in a single throw with two dice, the chance of throwing more than seven is equal to that of throwing less than seven.

Ans. 5/12 **[Delhi Univ. B.Sc., 1987, 1985]**

(*d*) In a single throw with two dice, what is the number whose probability is minimum?

(e) Two persons A and B throw three dice (six faced). If A throws 14, find B's chance of throwing a higher number. [Meerut Univ. B.Sc.(Stat.), 1987]

3. *(a)* A bag contains 7 white, 6 red and 5 black balls. Two balls are drawn at random. Find the probability that they will both be white.

Ans. 21/153

(b) A bag contains 10 white, 6 red, 4 black and 7 blue balls. 5 balls are drawn at random. What is the probability that 2 of them are red and one black?

Ans. ${}^6C_2 \times {}^4C_1 \Big/ {}^{27}C_5$

4. *(a)* From a set of raffle tickets numbered 1 to 100, three are drawn at random. What is the probability that all the tickets are odd-numbered?

Ans. ${}^{50}C_3 \Big/ {}^{100}C_3$

(b) A number is chosen from each of the two sets :

$$(1, 2, 3, 4, 5, 6, 7, 8, 9); \quad (4, 5, 6, 7, 8, 9)$$

If p_1 is the probability that the sum of the two numbers be 10 and p_2 the probability that their sum be 8, find $p_1 + p_2$.

(c) Two different digits are chosen at random from the set 1,2,3,...,8. Show that the probability that the sum of the digits will be equal to 5 is the same as the probability that their sum will exceed 13, each being 1/14. Also show that the chance of both digits exceeding 5 is 3/28. [Nagpur Univ. B.Sc., 1992]

5. What is the chance that *(i)* a leap year selected at random will contain 53 Sundays? *(ii)* a non-leap year selected at random would contain 53 Sundays.

Ans. (i) 2/7, (ii) 1/7

6. *(a)* What is the probability of having a knave and a queen when two cards are drawn from a pack of 52 cards?

Ans. 8/663

(b) Seven cards are drawn at random from a pack of 52 cards. What is the probability that 4 will be red and 3 black?

Ans. ${}^{26}C_4 \times {}^{26}C_3 \Big/ {}^{52}C_7$

(c) A card is drawn from an ordinary pack and a gambler bets that it is a spade or an ace. What are the odds against his winning the bet?

Ans. 9:4

(d) Two cards are drawn from a pack of 52 cards. What is the chance that

(i) they belong to the same suit?

(ii) they belong to different suits and different denominations.

[Bombay Univ. B.Sc., 1986]

7. *(a)* If the letters of the word RANDOM be arranged at random, what is the chance that there are exactly two letters between A and O.

(b) Find the probability that in a random arrangement of the leters of the word 'UNIVERSITY', the two I's do not come together.

(c) In random arrangements of the letters of the word 'ENGINEERING', what is the probability that vowels always occur together?

[Kurushetra Univ. B.E., 1991]

(d) Letters are drawn one at a time from a box containing the letters A, H, M, O, S, T. What is the probability that the letters in the order drawn spell the word 'Thomas'?

8. A letter is taken out at random out of 'ASSISTANT' and a letter out of 'STATISTIC'. What is the chance that they are the same letters?

9. *(a)* Twelve persons amongst whom are x and y sit down at random at a round table. What is the probability that there are two persons between x and y?

(b) A and B stand in a line at random with 10 other people. What is the probability that there will be 3 persons between A and B?

10. *(a)* If from a lot of 30 tickets marked 1, 2, 3, ..., 30 four tickets are drawn, what is the chance that those marked 1 and 2 are among them?

Ans. 2/145

(b) A bag contains 50 tickets numbered 1, 2, 3, ..., 50 of which five are drawn at random and arranged in ascending order of the magnitude ($x_1 < x_2 < x_3 < x_4 < x_5$). What is the probability that $x_3 = 30$?

Hint. *(a)* Exhaustive number of cases $= {}^{30}C_4$

If, of the four tickets drawn, two tickets bear the numbers 1 and 2, the remaining 2 must have come out of 28 tickets numbered from 3 to 30 and this can be done in ${}^{28}C_2$ ways.

$\therefore$ Favourable number of cases $= {}^{28}C_2$

(b) Exhaustive number of cases $= {}^{50}C_5$

If $x_3 = 30$, then the two tickets with numbers x_1 and x_2 must have come out of 29 tickets numbered from 1 to 29 and this can be done in ${}^{29}C_2$ ways, and the other two tickets with numbers x_4 and x_5 must have been drawn out of 20 tickets numbered from 31 to 50 and this can be done in ${}^{20}C_2$ ways.

$\therefore$ No. of favourable cases $= {}^{29}C_2 \times {}^{20}C_2$.

11. Four persons are chosen ar random from a group containing 3 men, 2 women and 4 children. Show that the chance that exactly two of them will be children is 10/21. **[Delhi Univ. B.A.1988]**

Ans. $\dfrac{{}^4C_2 \times {}^5C_2}{{}^9C_4} = \dfrac{10}{21}$

12. From a group of 3 Indians, 4 Pakistanis and 5 Americans a sub-committee of four people is selected by lots. Find the probability that the sub-committee will consist of

(i) 2 Indians and 2 Pakistanis

(ii) 1 Indian, 1 Pakistani and 2 Americans

(iii) 4 Americans [Madras Univ. B.Sc.(Main Stat.), 1987]

Ans. (*i*) $\dfrac{{}^3C_2 \times {}^4C_2}{{}^{12}C_4}$, (*ii*) $\dfrac{{}^3C_1 \times {}^4C_1 \times {}^5C_2}{{}^{12}C_4}$, (*iii*) $\dfrac{{}^5C_4}{{}^{12}C_4}$

13. In a box there are 4 granite stones, 5 sand stones and 6 bricks of identical size and shape. Out of them 3 are chosen at random. Find the chance that :

(i) They all belong to different varieties.

(ii) They all belong to the same variety.

(iii) They are all granite stones. **(Madras Univ. B.Sc., Oct. 1992)**

14. If n people are seated at a round table, what is the chance that two named individuals will be next to each other?

Ans. 2/(n-1)

15. Four tickets marked 00, 01, 10 and 11 respectively are placed in a bag. A ticket is drawn at random five times, being replaced each time. Find the probability that the sum of the numbers on tickets thus drawn is 23.

[Delhi Univ. B.Sc.(Subs.), 1988]

16. From a group of 25 persons, what is the probability that all 25 will have different birthdays? Assume a 365 day year and that all days are equally likely.

[Delhi Univ. B.Sc.(Maths Hons.), 1987]

Hint. $(365 \times 364 \times \ldots \times 341) \div (365)^{25}$

4·4. Mathematical Tools : Preliminary Notions of Sets. The set theory was developed by the German mathematician, G. Cantor (1845–1918).

4·4·1. Sets and Elements of Sets. A set is a well defined collection or aggregate of all possible objects having given properties and specified according to a well defined rule. The objects comprising a set are called elements, members or points of the set. Sets are often denoted by capital letters, *viz.*, A, B, C, etc. If x is an element of the set A, we write symbolically $x \in A$ (x belongs to A). If x is not a member of the set A, we write $x \notin A$ (x does not belong to A). Sets are often described by describing the properties possessed by their members. Thus the set A of all non-negative rational numbers with square less than 2 will be written as $A = \{x : x \text{ rational}, x \geq 0, x^2 < 2\}$.

If every element of the set A belongs to the set B, *i.e.*, if $x \in A \Rightarrow x \in B$, then we say that A is a subset of B and write symbolically $A \subseteq B$ (A is contained in B) or $B \supseteq A$ (B contains A). Two sets A and B are said to be *equal* or *identical* if $A \subseteq B$ and $B \subseteq A$ and we write $A = B$ or $B = A$.

A *null* or an *empty* set is one which does not contain any element at all and is denoted by ϕ.

Remarks. 1. Every set is a subset of itself.

2. An empty set is subset of every set.

3. A set containing only one element is conceptually distinct from the element itself, but will be represented by the same symbol for the sake of convenience.

4. As will be the case in all our applications of set theory, especially to probability theory, we shall have a fixed set S (say) given in advance, and we shall be concerned only with subsets of this given set. The underlying set S may vary from one application to another, and it will be referred to as *universal set* of each particular discourse.

4·4·2. Operation on Sets

The union of two given sets A and B, denoted by $A \cup B$, is defined as a set consisting of all those points which belong to either A or B or both. Thus symbolically,

$$A \cup B = \{ x : x \in A \text{ or } x \in B \}.$$

Similarly

$$\bigcup_{i=1}^{n} A_i = \{ x : x \in A_i \text{ for at least one } i = 1, 2, \ldots, n \}$$

The *intersection* of two sets A and B, denoted by $A \cap B$, is defined as a set consisting of all those elements which belong to both A and B. Thus

$$A \cap B = \{ x : x \in A \text{ and } x \in B \}.$$

Similarly

$$\bigcap_{i=1}^{n} A_i = \{ x : x \in A_i \text{ for all } i = 1, 2, \ldots, n \}$$

For example, if $A = \{1, 2, 5, 8, 10\}$ and $B = \{2, 4, 8, 12\}$, then

$A \cup B = \{1, 2, 4, 5, 8, 10, 12\}$ and $A \cap B = \{2, 8\}$.

If A and B have no common point, *i.e.*, $A \cap B = \phi$, then the sets A and B are said to be *disjoint, mutually exclusive* or *non-overlapping*.

The *relative difference* of a set A from another set B, denoted by $A-B$ is defined as a set consisting of those elements of A which do not belong to B. Symbolically,

$$A - B = \{ x : x \in A \text{ and } x \notin B \}.$$

The *complement* or *negative* of any set A, denoted by $\bar{A}$ is a set containing all elements of the universal set S, (say), that are not elements of A, *i.e.*, $\bar{A} = S - A$.

4·4·3. Algebra of Sets

Now we state certain important properties concerning operations on sets. If A, B and C are the subsets of a universal set S, then the following laws hold:

Commutative Law : $A \cup B = B \cup A$, $A \cap B = B \cap A$

Associative Law : $(A \cup B) \cup C = A \cup (B \cup C)$

$(A \cap B) \cap C = A \cap (B \cap C)$

Distributive Law : $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Complementary Law : $A \cup \bar{A} = S$, $A \cap \bar{A} = \phi$

$A \cup S = S$, $(\because A \subseteq S)$, $A \cap S = A$

$A \cup \phi = A$, $A \cap \phi = \phi$

Difference Law : $A - B = A \cap \bar{B}$

$A - B = A - (A \cap B) = (A \cup B) - B$

$A - (B - C) = (A - B) \cup (A - C)$.

$$(A \cup B) - C = (A - C) \cup (B - C)$$
$$A - (B \cup C) = (A - B) \cap (A - C)$$
$$(A \cap B) \cup (A - B) = A\,,\ (A \cap B) \cap (A - B) = \phi$$

De-Morgan's Law—Dualization Law.

$$\overline{(A \cup B)} = \bar{A} \cap \bar{B} \text{ and } \overline{(A \cap B)} = \bar{A} \cup \bar{B}$$

More generally

$$\overline{\left(\bigcup_{i=1}^{n} A_i\right)} = \bigcap_{i=1}^{n} \bar{A}_i \quad \text{and} \quad \overline{\left(\bigcap_{i=1}^{n} A_i\right)} = \bigcup_{i=1}^{n} \bar{A}_i$$

Involution Law : $\overline{(\bar{A})} = A$

Idempotency Law : $A \cup A = A\,,\quad A \cap A = A$

4·4·4. Limit of Sequence of Sets

Let $\{A_n\}$ be a sequence of sets in S. The *limit supremum* or *limit superior* of the sequence, usually written as lim sup A_n, is the set of all those elements which belong to A_n for infinitely many n. Thus

$$\limsup_{n \to \infty} A_n = \{ x : x \in A_n \text{ for infinitely many } n \} \qquad \text{...(4·3)}$$

The set of all those elements which belong to A_n for all but a finite number of n is called *limit infinimum or limit inferior* of the sequence and is denoted by lim inf A_n. Thus

$$\liminf_{n \to \infty} A_n = \{ x : x \in A_n \text{ for all but a finite number of } n \} \qquad \text{...(4·3 a)}$$

The sequence $\{A_n\}$ is said to have a limit if and only if lim sup A_n = lim inf A_n and this common value gives the limit of the sequence.

Theorem 4·1. *lim sup* $A_n = \bigcap_{m=1}^{\infty} \left(\bigcup_{n=m}^{\infty} A_n \right)$

and *lim inf* $A_n = \bigcup_{m=1}^{\infty} \left(\bigcap_{n=m}^{\infty} A_n \right)$

Def. $\{A_n\}$ is a monotone (infinite) sequence of sets if either

(i) $A_n \subset A_{n+1}\ \forall\ n$ or *(ii)* $A_n \supset A_{n+1}\ \forall\ n$.

In the former case the sequence $\{A_n\}$ is said to be *non-decreasing sequence* and is usually expressed as $A_n \uparrow$ and in the latter case it is said to be *non-increasing sequence* and is expressed as $A_n \downarrow$.

For a monotone sequence (non-increasing or non-decreasing), the limit always exists and we have,

$$\lim_{n \to \infty} A_n = \begin{cases} \bigcup_{n=1}^{\infty} A_n \text{ in case } (i)\,,\ i.e., A_n \uparrow \\ \bigcap_{n=1}^{\infty} A_n \text{ in case } (ii)\,,\ i.e., A_n \downarrow \end{cases}$$

4·4·5. Classes of Sets. A group of sets will be termed as a *class* (of sets). Below we shall define some useful types of classes.

A ring **R** is a *non-empty* class of sets which is closed under the formation of 'finite unions' and 'difference',

i.e., if $A \in \mathbf{R}$, $B \in \mathbf{R}$, then $A \cup B \in \mathbf{R}$ and $A - B \in \mathbf{R}$.

Obviously ϕ is a member of every ring.

A *field* **F** (or an *algebra*) is a non-empty class of sets which is closed under the formation of finite unions and under complementation. Thus

(i) $A \in \mathbf{F}, B \in \mathbf{F} \Rightarrow A \cup B \in \mathbf{F}$ and

(ii) $A \in \mathbf{F} \Rightarrow \bar{A} \in \mathbf{F}$.

A *σ-ring* **C** is a non-empty class of sets which is closed under the formation of 'countable unions' and 'difference'. Thus

(i) $A_i \in \mathbf{C}, i = 1, 2, \ldots \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathbf{C}$

(ii) $A \in \mathbf{C}, B \in \mathbf{C} \quad \Rightarrow \quad A - B \in \mathbf{C}$.

More precisely σ-ring is a ring which is closed under the formation of countable unions.

A σ field (or σ-algebra) **B** is a non-empty class of sets that is closed under the formation of 'countable unions' and complementations,

i.e.,

(i) $A_i \in \mathbf{B}, i = 1, 2, \ldots \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathbf{B}$.

(ii) $A \in \mathbf{B} \quad \Rightarrow \quad \bar{A} \in \mathbf{B}$.

σ-field may also be defined as a field which is closed under the formation of countable unions.

4·5. Axiomatic Approach to Probability. The axiomatic approach to probability, which closely relates the theory of probability with the modern metric theory of functions and also set theory, was proposed by A.N. Kolmogorov, a Russian mathematician, in 1933. The axiomatic definition of probability includes 'both' the classical and the statistical definitions as particular cases and overcomes the deficiencies of each of them. On this basis, it is possible to construct a logically perfect structure of the modern theory of probability and at the same time to satisfy the enchanced requirements of modern natural science. The axiomatic development of mathematical theory of probability relies entirely upon the logic of deduction.

The diverse theorems of probability, as were available prior to 1933, were finally brought together into a unified axiomised system in 1933. It is important to remark that probability theory, in common with all axiomatic mathematical systems, is concerned solely with relations among undefined things.

The axioms thus provide a set of rules which define relationships between abstract entities. These rules can be used to deduce theorems, and the theorems can be brought together to deduce more complex theorems. These theorems have no empirical meaning although they can be given an interpretation in terms of empirical phenomenon. The important point, however, is that the mathematical development of probability theory is, in no way, conditional upon the interpretation given to the theory.

More precisely, under axiomatic approach, the probability can be deduced from mathematical concepts. To start with some concepts are laid down. Then some statements are made in respect of the properties possessed by these concepts. These properties, often termed as "*axioms*" of the theory, are used to frame some theorems. These theorems are framed without any reference to the real world and are deductions from the axioms of the theory.

4·5·1. Random Experiment, Sample Space. The theory of probability provides *mathematical models* for "real-world phenomenon" involving games of chance such as the tossing of coins and dice. We feel intuitively that statements such as

(i) "The probability of getting a "head" in one toss of an unbiased coin is 1/2"

(ii) "The probability of getting a "four" in a single toss of an unbiased die is 1/6",

should hold. We also feel that the probability of obtaining *either* a "5" or a "6" in a single throw of a die, should be the sum of the probabilities of a "5" and a "6", *viz.*, 1/6+1/6=1/3. That is, probabilities should have some kind of *additive* property. Finally, we feel that in a large number of repetitions of, for example, a coin tossing experiment, the proportion of heads should be approximately 1/2. That is, the probability should have a *frequency interpretation.*

To deal with these properties sensibly, we need a *mathematical description* or *model* for the probabilistic situation we have. Any such probabilistic situation is referred to as a *random experiment*, denoted by **E**. **E** may be a coin or die throwing experiment, drawing of cards from a well-shuffled pack of cards, an agricultural experiment to determine the effects of fertilizers on yield of a commodity, and so on.

Each performance in a random experiment is called a *trial*. That is, all the trials conducted under the same set of conditions form a random experiment. The result of a trial in a random experiment is called an outcome, an elementary event or a sample point. The totality of all possible outcomes (*i.e.*, sample points) of a random experiment constitutes the *sample space*.

Suppose $e_1, e_2, ..., e_n$ are the possible outcomes of a random experiment E such that no two or more of them can occur simultaneously and exactly one of the outcomes $e_1, e_2, ..., e_n$ must occur. More specifically, with an experiment E, we associated a set $S = (e_1, e_2, ..., e_n)$ of possible outcomes with the following properties:

(i) each element of S denotes a possible outcome of the experiement,

(ii) any trial results in an outcome that corresponds to one and only one element of the set S.

The set S associated with an experiment E, real or conceptual, satisfying the above two properties is called the *sample space* of the experiment.

Remarks. 1. The sample space serves as universal set for all questions concerned with the experiment.

2. A sample space S is said to be finite (infinite) sample sapce if the number of elements in S is finite (infinite). For example, the sample space associated with the experiment of throwing the coin until a head appears, is infinite, with possible *sample points*

$$\{\omega_1,\ \omega_2,\ \omega_3,\ \omega_4,\ ....\}$$

where $\omega_1 = H,\ \omega_2 = TH,\ \omega_3 = TTH,\ \omega_4 = TTTH$, and so on, H denoting a head and T a tail.

3. A sample space is called discrete if it contains only finitely or infinitely many points which can be arranged into a simple sequence $\omega_1,\ \omega_2,\ ...,$ while a sample space containing non-denumerable number of points is called a continuous sample space. In this book, we shall restrict ourselves to discrete sample spaces only.

4·5·2. Event. Every non-empty subset A of S, which is a disjoint union of single element subsets of the sample space S of a random experiment E is called an event. The notion of an event may also be defined as follows:

"Of all the possible outcomes in the sample space of an experiment,some outcomes satisfy a specified description,which we call an event."

Remarks. 1. As the empty set ϕ is a subset of S , ϕ is also an event, known as *impossible* event.

2. An event A, in particular, can be a single element subset of S, in which case it is known as *elementary* event.

4·5·3. Some Illustrations — Examples. We discuss below some examples to illustrate the concepts of sample space and event.

1. Consider tossing of a coin singly. The possible outcomes for this experiment are (writing H for a "head" and T for a "tail") : H and T. Thus the sample space S consists of two points $\{\omega_1,\ \omega_2\}$, corresponding to each possible outcome or elementary event listed.

i.e., $\quad S = \{\omega_1,\ \omega_2\} = \{H,T\}$ and $n(S) = 2$,

where $n(S)$ is the total number of sample points in S.

If we consider two tosses of a coin, the possible outcomes are HH, HT, TH, TT. Thus, in this case the sample space S consists of four points $\{\omega_1, \omega_2, \omega_3, \omega_4\}$, corresponding to each possible outcome listed and $n(S) = 4$. Combinations of these outcomes form what we call events. For example, the event of getting at least one head is the set of the outcomes $\{HH,HT,TH\} = \{\omega_1, \omega_2, \omega_3\}$. Thus, mathematically, the events are subsets of S.

2. Let us consider a single toss of a die. Since there are six possible outcomes, our sample space S is now a space of six points $\{\omega_1, \omega_2, \ldots, \omega_6\}$ where ω_i corresponds to the appearance of number i. Thus $S=\{\omega_1, \omega_2, \ldots, \omega_6\}=\{1,2,\ldots,6\}$ and $n(S)=6$. The event that the outcome is even is represented by the set of points $\{\omega_2, \omega_4, \omega_6\}$.

3. A coin and a die are tossed together. For this experiment, our sample space consists of twelve points $\{\omega_1, \omega_2, \ldots, \omega_{12}\}$ where ω_i $(i = 1, 2, \ldots, 6)$ represents a head on coin together with appearance of ith number on the die and ω_i $(i = 7, 8, \ldots, 12)$ represents a tail on coin together with the appearance of ith number on die. Thus

$$S = \{\omega_1, \omega_2, \ldots, \omega_{12}\} = \{(H, T) \times (1, 2, \ldots, 6)\} \text{ and } n(S) = 12$$

Remark. If the coin and die are unbiased, we can see intuitively that in each of the above examples, the outcomes (sample points) are equally likely to occur.

4. Consider an experiment in which two balls are drawn one by one from an urn containing 2 white and 4 blue balls such that when the second ball is drawn, the first is *not* replaced.

Let us number the six balls as 1, 2, 3, 4, 5 and 6, numbers 1 and 2 representing a white ball and numbers 3, 4, 5, and 6 representing a blue ball. Suppose in a draw we pick up balls numbered 2 and 6. Then (2,6) is called an outcome of the experiment. It should be noted that the outcome (2,6) is different from the outcome (6,2) because in the former case ball No. 2 is drawn first and ball No.6 is drawn next while in the latter case, 6th ball is drawn first and the second ball is drawn next.

The sample space consists of thirty points as listed below:

$\omega_1 = (1,2)$	$\omega_2 = (1,3)$	$\omega_3 = (1,4)$	$\omega_4 = (1,5)$	$\omega_5 = (1,6)$
$\omega_6 = (2,1)$	$\omega_7 = (2,3)$	$\omega_8 = (2,4)$	$\omega_9 = (2,5)$	$\omega_{10} = (2,6)$
$\omega_{11} = (3,1)$	$\omega_{12} = (3,2)$	$\omega_{13} = (3,4)$	$\omega_{14} = (3,5)$	$\omega_{15} = (3,6)$
$\omega_{16} = (4,1)$	$\omega_{17} = (4,2)$	$\omega_{18} = (4,3)$	$\omega_{19} = (4,5)$	$\omega_{20} = (4,6)$,
$\omega_{21} = (5,1)$	$\omega_{22} = (5,2)$	$\omega_{23} = (5,3)$	$\omega_{24} = (5,4)$	$\omega_{25} = (5,6)$
$\omega_{26} = (6,1)$	$\omega_{27} = (6,2)$	$\omega_{28} = (6,3)$	$\omega_{29} = (6,4)$	$\omega_{30} = (6,5)$

Thus

$$S = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_{30}\} \text{ and } n(S) = 30$$

$$\Rightarrow \quad S = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\} - \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$$

The event

(i) the first ball drawn is white

(ii) the second ball drawn is white

(iii) both the balls drawn are white

(iv) both the balls drawn are black

are represented respectively by the following sets of points:

$\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8, \omega_9, \omega_{10}\}$,

$\{\omega_1, \omega_6, \omega_{11}, \omega_{12}, \omega_{16}, \omega_{17}, \omega_{21}, \omega_{22}, \omega_{26}, \omega_{27}\}$,

$\{\omega_1, \omega_6\}$, and

$\{\omega_{13}, \omega_{14}, \omega_{15}, \omega_{18}, \omega_{19}, \omega_{20}, \omega_{23}, \omega_{24}, \omega_{25}, \omega_{28}, \omega_{29}, \omega_{30}\}$.

5. Consider an experiment in which two dice are tossed. The sample space S for this experiment is given by

$$S = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$$

and $\quad n(S) = 6 \times 6 = 36.$

Let E_1 be the event that 'the sum of the spots on the dice is greater than 12', E_2 be the event that 'the sum of spots on the dice is divisible by 3', and E_3 be the event that 'the sum is greater than or equal to two and is less than or equal to 12'. Then these events are represented by the following subsets of S :

$$E_1 = \{\phi\},\ E_3 = S \text{ and}$$
$$E_2 = \{(1, 2), (1, 5), (2, 1), (2, 4), (3, 3), (3, 6), (4, 2), (4, 5), (5, 1), (5, 4), (6, 3), (6, 6)\}$$

Thus $\quad n(E_1) = 0,\ n(E_2) = 12,$ and $n(E_3) = 36$

Here E is an 'impossible event' and E_3 a 'certain event'.

6. Let E denote the experiment of tossing a coin three times in succession or tossing three coins at a time. Then the sample space S is given by

$$\begin{aligned} S &= \{H, T\} \times \{H, T\} \times \{H, T\} \\ &= \{H, T\} \times \{HH, HT, TH, TT\} \\ &= \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \\ &= \{\omega_1, \omega_2, \omega_3, \ldots, \omega_8\}, \text{ say.} \end{aligned}$$

If E_1 is the event that 'the number of heads exceeds the number of tails', E_2, the event of 'getting two heads' and E_3, the event of getting 'head in the first trial' then these are represented by the following sets of points :

$$E_1 = \{\omega_1, \omega_2, \omega_3, \omega_5\},$$
$$E_2 = \{\omega_2, \omega_3, \omega_5\}$$

and $\quad E_3 = \{\omega_1, \omega_2, \omega_3, \omega_4\}$.

7. In the foregoing examples the sample sapce is finite. To construct an experiment in which the sample sapce is countably infinite, we toss a coin repeatedly until head or tail appears twice in succession. The sample space of all the possible outcomes may be represented as :

$$S = \{HH, TT, THH, HTT, HTHH, THTT, THTHH, HTHTT, \ldots\}.$$

4·5·4. Algebra of Events. For events A, B, C

(i) $A \cup B = \{\omega \in S : \omega \in A \text{ or } \omega \in B\}$

(ii) $A \cap B = \{\omega \in S : \omega \in A \text{ and } \omega \in B\}$

(iii) $\bar{A}$ (*A complement*) $= \{\omega \in S : \omega \notin A\}$

(iv) $A - B = \{\omega \in S : \omega \in A \text{ but } \omega \notin B\}$

(v) Similar generalisations for $\bigcup_{i=1}^{n} A_i,\ \bigcap_{i=1}^{n} A_i,\ \bigcup_{i} A_i$ etc.

(vi) $A \subset B \Rightarrow$ for every $\omega \in A,\ \omega \in B$.

(vii) $B \supset A \Rightarrow A \subset B$.

(viii) $A = B$ if and only if A and B have the same elements, *i.e.*, if $A \subset B$ and $B \subset A$.

(ix) A and B disjoint (mutually exclusive) $\Rightarrow A \cap B = \phi$ (*null set*).

(x) $A \cup B$ can be denoted by $A + B$ if A and B are disjoint.

(xi) $A \Delta B$ denotes those ω belonging to exactly one of A and B, *i.e.*,

$$A \Delta B = A\overline{B} \cup \overline{A}B$$

Remark. Since the events are subsets of S, all the laws of set theory *viz.*, commutative laws, associative laws, distributive laws, De-Morgan's law, etc., hold for algebra of events.

Table – ***Glossary of Probability Terms***

	Statement	*Meaning in terms of set theory*
1.	At least one of the events A or B occurs.	$\omega \in A \cup B$
2.	Both the events A and B occur.	$\omega \in A \cap B$
3.	Neither A nor B occurs	$\omega \in \overline{A} \cap \overline{B}$
4.	Event A occurs and B does not occur	$\omega \in A \cap \overline{B}$
5.	Exactly one of the events A or B occurs.	$\omega \in A \Delta B$
6.	Not more than one of the events A or B occurs.	$\omega \in (A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (\overline{A} \cap \overline{B})$
7.	If event A occurs, so does B	$A \subset B$
8.	Events A and B are mutually exclusive.	$A \cap B = \phi$
9.	Complementary event of A.	$\overline{A}$
10.	Sample space	universal set S

Example 4·11. *A, B and C are three arbitrary events. Find expressions for the events noted below, in the context of A, B and C.*

(i) only A occurs,

(ii) Both A and B, but not C, occur,

(iii) All three events occur,

(iv) At least one occurs,

(v) At least two occur,

(vi) One and no more occurs,

(vii) Two and no more occur,

(viii) None occurs.

Solution.

(i) $A \cap \overline{B} \cap \overline{C}$, *(ii)* $A \cap B \cap \overline{C}$, *(iii)* $A \cap B \cap C$,

(iv) $A \cup B \cup C$,

(v) $(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C) \cup (A \cap B \cap C)$

(vi) $(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)$

(vii) $(A \cap B \cap \bar{C}) \cup (\bar{A} \cap B \cap C) \cup (A \cap \bar{B} \cap C)$

(viii) $\bar{A} \cap \bar{B} \cap \bar{C}$ or $\overline{A \cup B \cup C}$

EXERCISE 4(b)

1. *(i)* If A, B and C are any three events, write down the theoretical expressions for the following events:

(a) Only A occurs, *(b)* A and B occur but C does not,

(c) A, B, and C all the three occur, *(d)* at least one occurs

(e) At least two occur, *(f)* one does not occur,

(g) Two do not occurs, and *(h)* None occurs.

(ii) A, B and C are three events. Express the following events in appropriate symbols:

(a) Simultaneous occurrence of A, B and C.

(b) C currence of at least one of them.

(c) A, B and C are mutually exclusive events.

(d) Every point of A is contained in B.

(e) The event B but not A occurs. **[Gauhati Univ. B.Sc., Oct.1990]**

2. A sample space S contains four points x_1, x_2, x_3 and x_4 and the values of a set function $P(A)$ are known for the following sets :

$$A_1 = (x_1, x_2) \text{ and } P(A_1) = \frac{4}{10};\ A_2 = (x_3, x_4) \text{ and } P(A_2) = \frac{6}{10};$$

$$A_3 = (x_1, x_2, x_3) \text{ and } P(A_3) = \frac{4}{10};\ A_4 = (x_2, x_3, x_4) \text{ and } P(A_4) = \frac{7}{10}$$

Show that :

(i) the total number of sets (including the "null" set of number points) of points of x is 16.

(ii) Although the set containing no sample point has zero probability, the converse is not always true, *i.e.*, a set may have zero probability and yet it may be the set of a number of points.

3. Describe explicitly the sample spaces for each of the following experiments:

(i) The tossing of four coins.

(ii) The throwing of three dice.

(iii) The tossing of ten coins with the aim of observing the numbers of tails coming up.

(iv) Two cards are selected from a standard deck of cards.

(v) Four successive draws *(a)* with replacement, and *(b)* without replacement, from a bag containing fifty coloured balls out of which ten are white, twenty blue and twenty red.

(vi) A survey of families with two children is conduced and the sex of the children (the older child first) is recorded.

(vii) A survey of families with three children is made and the sex of the children (in order of age, oldest child first) are recorded.

(*viii*) Three distinguishable objects are distributed in three numbered cells.

(*ix*) A poker hand (five cards) is dealt from an ordinary deck of cards.

(*x*) Selecting r screws from the lot produced by a machine, a screw can be defective or non-defective.

4. In an experiment a coin is thrown five times. Write down the sample space. How many points are there in the sample space?

5. Describe sample space appropriate in each of the following cases :

(*i*) n-tosses of a coin with head or tails as outcome in each toss.

(*ii*) Successive tosses of a coin until a head turns up.

(*iii*) A survey of families with two children is conducted and the sex of the children (the older child first) is recorded.

(*iv*) Two successive draws, (*a*) with replacement (*b*) without replacement, from a bag containing 4 coloured toys out of which one is white, one black and 2 red toys. **[M.S.Baroda Univ. B.Sc., 1991]**

6. (*a*) An experiment consists of tossing an unbiased coin until the same result appears twice on succession for the first time. To every possible outcome requiring n tosses attribute probability $1/2^n$. Describe the sample space.

(*b*) A coin is tossed until there are either two consecutive heads or two consecutive tails or the number of tosses becomes five. Describe the sample space along with the probability associated with each sample point, if every sequence of n tosses has probabilty 2^{-n}. **[Civil Services (main), 1983]**

7. Urn 1 contains two white, one red and 3 black balls. Urn 2 contains one white, 3 red and 2 black balls. An experiment consists of first selecting an urn and then drawing a ball from this urn. Define a suitable sample space for this experiment.

8. Suppose an experiment has n outcomes $A_1, A_2, ..., A_n$ and that it is repeated r times. Let $x_1, x_2, ..., x_n$ record the number of occurrences of $A_1, A_2, ..., A_n$. Describe the sample space. Show that the number of sample points is

$$\binom{n+r-1}{r-1}$$

9. A manufacturer buys parts from four different vendors numbered 1, 2, 3 and 4. Referring to orders placed on two successive days, (1,4) denotes the event that on the first day, the order was given to vendor 1 and on the second day it was given to vendor 4. Letting A represent the event that vendor 1 gets at least one of these two orders, B the event that the same vendor gets both orders and C the event that vendors 1 and 3 do not get either order. List the elements of :

(*a*) entire sample space, (*b*) A, (*c*) B, (*d*) C, (*e*) $\bar{A}$, (*f*) $\bar{B}$, (*g*) $B \cup C$, (*h*) $A \cap B$, (*i*) $A \cap C$, (*j*) $\overline{A \cup B}$, and (*k*) $A - B$

[**Hint.** (*a*) The elements of entire sample space are

(1,1); (1, 2); (1, 3); (1, 4); (2, 1); (2, 2); (2, 3); (2, 4);
(3, 1); (3, 2); (3, 3); (3, 4); (4, 1); (4, 2); (4, 3); (4, 4).

(b) The elements of A are

$(1, 1); (1, 2); (1, 3); (1, 4); (2, 1); (3, 1); (4, 1)$.

(c) The elements of B are $(1, 1); (2, 2); (3, 3)$ and $(4, 4)$.

(d) The elements of C are $(2, 2); (2, 4); (4, 2); (4, 4)$.

(e) The element of $\overline{A}$ are :

$(2, 2); (2, 3); (2, 4); (3, 2); (3, 3); (3, 4); (4, 2); (4, 3); (4, 4)$.

(f) The elements of $\overline{B}$ are : $(1, 2); (1, 3); (1, 4); (2, 1); (2, 3); (2, 4); (3, 1); (3, 2); (3, 4); (4, 1); (4, 2); (4, 3)$.

(g) The elements of $B \cup C$ are $(1, 1); (2, 2); (3, 3); (4, 4); (2, 4); (4, 2)$.

(h) The elements of $A \cap B$ are $(1, 1)$.

(i) $A \cap C = \phi$

(j) Since $\overline{A \cup B} = \overline{A} \cap \overline{B}$. The elements of $\overline{A \cup B}$ are $(2, 3); (2, 4); (3, 2); (3, 4); (4, 2); (4, 3)$.

(k) The elements of $A - B$ are $(1, 2); (1, 3); (1, 4); (2, 1); (3, 1); (4, 1)$.

4·6. Probability — Mathematical Notion. We are now set to give the mathematical notion of the occurrence of a random phenomenon and the mathematical notion of probability. Suppose in a large number of trials the sample space S contains N sample points. The event A is defined by a description which is satisfied by N_A of the occurrences. The frequency interpretation of the probability $P(A)$ of the event A, tells us that $P(A) = N_A/N$.

A purely mathematical definition of probability cannot give us the actual value of $P(A)$ and this must be considered as a function defined on all events. With this in view, a mathematical definition of probability is enunciated as follows:

"Given a sample description space, probability is a function which assigns a non-negative real number to every event A, denoted by P(A) and is called the probability of the event A."

4·6·1. Probability Function. $P(A)$ is the probability function defined on a σ-field **B** of events if the following properties or axioms hold :

1. For each $A \in$ **B**, $P(A)$ is defined, is real and $P(A) \geq 0$

2. $P(S) = 1$

3. If $\{A_n\}$ is any finite or infinite sequence of disjoint events in **B**, then

$$P\left(\bigcup_{i=1}^{n} A_n\right) = \sum_{i=1}^{n} P(A_i) \qquad \ldots(4\cdot4)$$

The above three axioms are termed as the axiom of positiveness, certainty and union (additivity), respectively.

Remarks. 1. The set function **P** defined on σ-field **B**, taking its values in the real line and satisfying the above three axioms is called the probability measure.

2. The same definition of probability applies to *uncountable sample space* except that special restrictions must be placed on S and its subsets. It is important to realise that for a complete description of a probability measure, three things must be specified, *viz.*, the sample space S, the σ-field (σ-algebra) **B** formed from certain subset of S and set function **P**. The triplet (**S**, **B**, **P**,) is often called the *probability space*. In most elementary applications, S is finite and the σ-algebra **B** is taken to be the collection of all subsets of S.

3. It is interesting to see that there are some formal statements of the properties of events derived from the frequency approach. Since $P(A)=N_A/N$, it is easy to see that $P(A) \geq 0$, as in Axiom 1. Next since $N_S=N$, $P(S)=1$, as in Axiom 2. In case of two mutually exclusive (or disjoint) events A and B defined by sample points N_A and N_B, the sample points belonging to $A \cup B$ are N_A+N_B. Therefore,

$$P(A \cup B)=\frac{N_A+N_B}{N}=\frac{N_A}{N}+\frac{N_B}{N}=P(A)+P(B), \text{ as in axiom 3.}$$

Extended Axiom of Addition. If an event A can materialise in the occurrence of any one of the pairwise disjoint events A_1, A_2, ... so that

$$A=\bigcup_{i=1}^{\infty} A_i\,;\; A_i \cap A_j=\phi \quad (i \neq j)$$

then

$$P(A)=P\left(\bigcup_{i=1}^{\infty} A_i\right)=\sum_{i=1}^{\infty} P(A_i) \qquad \text{...(1)}$$

Axiom of Continuity. If $B_1, B_2, \ldots, B_n, \ldots$ be a countable sequences of events such that

$$(i)\; B_i \supset B_{i+1}, \quad (i=1, 2, 3, \ldots)$$

and

$$(ii) \bigcap_{n=1}^{\infty} B_n=\phi$$

i.e., if each succeeding event implies the preceeding event and if their simultaneous occurrence is an impossible event then

$$\lim_{n \to \infty} P(B_n)=0 \qquad \text{...(2)}$$

We shall now prove that these two axioms, *viz.*, the extended axiom of addition and axiom of continuity are equivalent, *i.e.*, each implies the other, *i.e.*, (1) $\Leftrightarrow$ (2).

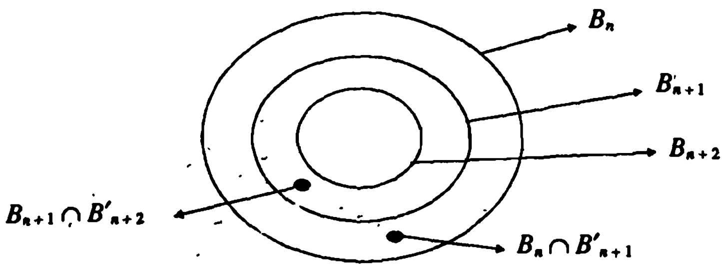
Theorem 4·1. *Axiom of continuity follows from the extended axiom of addition and vice versa.*

Proof. *(a)* (1) $\Rightarrow$ (2). Let $\{B_n\}$ be a countable sequence of events such that

$$B_1 \supset B_2 \supset B_3 \supset \ldots, \supset B_n \supset B_{n+1} \supset \ldots$$

and let for any $n \geq 1$,

$$\bigcap_{k \geq n} B_k=\phi \qquad (*)$$



Then it is obvious from the diagram that

$$B_n = B_n B'_{n+1} \cup B_{n+1} B'_{n+2} \cup \dots \cup \left(\bigcap_{k \geq n} B_k \right)$$

$$\Rightarrow \qquad B_n = \left(\bigcup_{k=n}^{\infty} B_k B'_{k+1} \right) \cup \left(\bigcap_{k \geq n} B_k \right),$$

where the events $B_k B'_{k+1}$; $(k=n, n+1, \dots)$ are pairwise disjoint and each is disjoint with $\bigcap_{k \geq n} B_k$.

Thus B_n has been expressed as the countable union of pairwise disjoint events and hence by the extended axiom of addition, we get

$$P(B_n) = \sum_{k=n}^{\infty} P(B_k B'_{k+1}) + P\left(\bigcap_{k \geq n} B_k \right)$$

$$= \sum_{k=n}^{\infty} P(B_k B'_{k+1}), \qquad (**)$$

since, from (*)

$$P\left(\bigcap_{k \geq n} B_k \right) = P(\phi) = 0$$

Further, from (**), since

$$\sum_{k=1}^{\infty} P(B_k B'_{k+1}) = P(B_1) \leq 1,$$

the right hand sum in (**), being the remainder after n terms of a convergent series tends to zero as $n \to \infty$.

Hence

$$\lim_{n \to \infty} P(B_n) = \lim_{n \to \infty} \sum_{k=n}^{\infty} P(B_k B'_{k+1}) = 0$$

Thus the extended axiom of addition implies the axiom of continuity.

(b) **Conversely (2) $\Rightarrow$ (1),** *i.e., the extended axiom of addition follows from the axiom of continuity.*

Let $\{A_n\}$ be a countable sequence of pairwise disjoint events and let

$$A = \bigcup_{i=1}^{\infty} A_i$$

$$= \left(\bigcup_{i=1}^{n} A_i\right) \cup \left(\bigcup_{i=n+1}^{\infty} A_i\right) \quad \text{...(3)}$$

Let us define a countable sequence $\{B_n\}$ of events by

$$B_n = \bigcup_{i=n}^{\infty} A_i \quad \text{...(4)}$$

Obviously B_n is a decreasing sequence of events, *i.e.*,

$$B_1 \supset B_2 \supset \ldots \supset B_n \supset B_{n+1} \supset \ldots. \quad \text{...(5)}$$

Also we have

$$A = \left(\bigcup_{i=1}^{n} A_i\right) \cup B_{n+1} \quad \text{...(6)}$$

Since A_i 's are pairwise disjoint, we get

$$A_i \cap B_{n+1} = \phi, \ (i = 1, 2, \ldots, n) \quad \text{...(6a)}$$

From (4) we see that if the event B_n has occurred it implies the occurrence of any one of the events $A_{n+1}, A_{n+2}, \ldots$ Without loss of generality let us assume that this event is A_i $(i = n+1, n+2, \ldots)$. Further since A_i 's are pairwise disjoint, the occurrence of A_i implies that events $A_{i+1}, A_{i+2}, \ldots$ do not occur leading to the conclusion that $B_{i+1}, B_{i+2}, \ldots$ will not occur.

$$\Rightarrow \quad \bigcap_{i=n}^{\infty} B_i = \phi \quad \text{...(7)}$$

From (5) and (7), we observe that both the conditions of axiom of continuity are satisfied and hence we get

$$\lim_{n \to \infty} P(B_n) = 0 \quad \text{...(8)}$$

From (6), we get

$$P(A) = P\left[\left(\bigcup_{i=1}^{n} A_i\right) \cup B_{n+1}\right]$$

$$= \sum_{i=1}^{n} P(A_i) + P(B_{n+1})$$

(By axiom of Additivity)

$$\Rightarrow \quad P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \to \infty} \sum_{i=1}^{n} P(A_i) + \lim_{n \to \infty} (B_{n+1})$$

$$= \sum_{i=1}^{\infty} P(A_i), \qquad \text{[From (8)]}$$

which is the extended axiom of addition.

THEOREMS ON PROBABILITIES OF EVENTS

Theorem 4·2. *Probability of the impossible event is zero, i.e., $P(\phi) = 0$.*

Proof. Impossible event contains no sample point and hence the certain event S and the impossible event ϕ are mutually exclusive.

Hence $S \cup \phi = S$

$\therefore$ $P(S \cup \phi) = P(S)$

$\Rightarrow$ $P(S) + P(\phi) = P(S)$ [By Axiom 3]

$\Rightarrow$ $P(\phi) = 0$

Remark. It may be noted $P(A)=0$, does not imply that A is necessarily an empty set. In practice, probability '0' is assigned to the events which are so rare that they happen only once in a lifetime. For example, if a person who does not know typing is asked to type the manuscript of a book, the probability of the event that he will type it correctly without any mistake is 0.

As another illustration, let us consider the random tossing of a coin. The event that the coin will stand erect on its edge, is assigned the probability 0.

The study of continuous random variable provides another illustration to the fact that $P(A)=0$, does not imply $A=\phi$, because in case of continous random variable X, the proability at a point is always zero, *i.e.*, $P(X=c)=0$ [See Chapter 5].

Theorem 4·3. Probability of the complementary event $\bar{A}$ of A is given by

$$P(\bar{A}) = 1 - P(A)$$

Proof. A and $\bar{A}$ are disjoint events.

Moreover, $A \cup \bar{A} = S$

From axioms 2 and 3 of probability, we have

$$P(A \cup \bar{A}) = P(A) + P(\bar{A}) = P(S) = 1$$

$\Rightarrow$ $P(\bar{A}) = 1 - P(A)$

Cor. 1. We have $P(A) = 1 - P(\bar{A})$

$\Rightarrow$ $P(A) \le 1$ $(\because P(\bar{A}) \ge 0)$

Cor. 2. $P(\phi) = 0$, since $\phi = \bar{S}$

and $P(\phi) = P(\bar{S}) = 1 - P(S) = 1 - 1 = 0.$

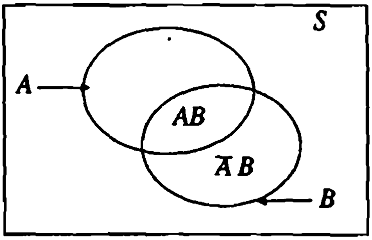
Theorem 4·4. *For any two events A and B,*

$$P(\bar{A} \cap B) = P(B) - P(A \cap B) \qquad \textbf{[Mysore Univ. B.Sc., 1992]}$$

Proof.

$\bar{A} \cap B$ and $A \cap B$ are disjoint events and

$(A \cap B) \cup (\bar{A} \cap B) = B$



Hence by axiom 3, we get

$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

$$\Rightarrow P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

Remark. Similarly, we shall get

$$P(A \cap \bar{B}) = P(A) - P(A \cap B)$$

Theorem 4·5. *Probability of the union of any two events A and B is given by*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof. $A \cup B$ can be written as the union of the two mutually disjoint events, A and $B \cap \bar{A}$.

$$\therefore \quad P(A \cup B) = P[A \cup (B \cap \bar{A})] = P(A) + P(B \cap \bar{A})$$

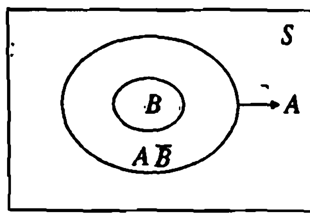
$$= P(A) + P(B) - P(A \cap B) \qquad (c.f. \text{ Theorem 4·4})$$

Theorem 4·6. *If $B \subset A$, then*

(i) $P(A \cap \bar{B}) = P(A) - P(B)$,

(ii) $P(B) \le P(A)$

Proof. *(i)* When $B \subset A$, B and $A \cap \bar{B}$ are mutually exclusive events and their union is A



Therefore

$$P(A) = P[B \cup (A \cap \bar{B})]$$

$$= P(B) + P(A \cap \bar{B}) \qquad \text{[By axiom 3]}$$

$$\Rightarrow \quad P(A \cap \bar{B}) = P(A) - P(B)$$

(ii) Using axiom 1,

$$P(A \cap \bar{B}) \ge 0 \quad \Rightarrow \quad P(A) - P(B) \ge 0$$

Hence $\quad P(B) \le P(A)$

Cor. Since $(A \cap B) \subset A$ and $(A \cap B) \subset B$,

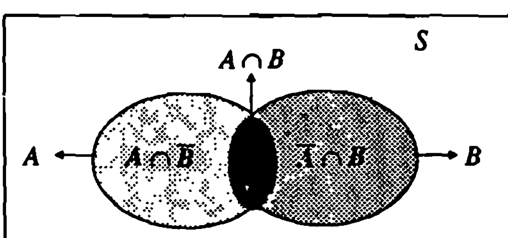
$$P(A \cap B) \le P(A) \quad \text{and} \quad P(A \cap B) \le P(B)$$

4·6·2. Law of Addition of Probabilities

Statement. *If A and B are any two events [subsets of sample space S] and are not disjoint, then*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad ...(4·5)$$

Proof.



We have

$$A \cup B = A \cup (\bar{A} \cap B)$$

Since A and $(\bar{A} \cap B)$ are disjoint,

$$\begin{aligned} P(A \cup B) &= P(A) + P(\bar{A} \cap B) \\ &= P(A) + [P(\bar{A} \cap B) + P(A \cap B)] - P(A \cap B) \\ &= P(A) + P[(\bar{A} \cap B) \cup (A \cap B)] - P(A \cap B) \end{aligned}$$

[$\because$ $(\bar{A} \cap B)$ and $(A \cap B)$ *are disjoint*]

$$\Rightarrow \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Remark. An alternative proof is provided by Theorems 4·4 *and* 4·5.

4·6·3. Extention of General Law of Addition of Probabilities. *For n events $A_1, A_2, \ldots, A_n$, we have*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum\sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum\sum\sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k)$$
$$- \ldots + (-1)^{n-1} P(A_1 \cap A_2 \cap \ldots \cap A_n) \qquad \ldots(4\text{·}6)$$

Proof. For two events A_1 and A_2, we have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \qquad \ldots(*)$$

Hence (4·6) is true for $n = 2$.

Let us now suppose that (4·6) is true for $n = r$, (say). Then

$$P\left(\bigcup_{i=1}^{r} A_i\right) = \sum_{i=1}^{r} P(A_i) - \sum\sum_{1 \le i < j \le r} P(A_i \cap A_j) + \ldots + (-1)^{r-1} P(A_1 \cap A_2 \cap \ldots \cap A_r) \qquad \ldots(**)$$

Now

$$P\left(\bigcup_{i=1}^{r+1} A_i\right) = P\left[\left(\bigcup_{i=1}^{r} A_i\right) \cup A_{r+1}\right]$$

$$= P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) - P\left[\left(\bigcup_{i=1}^{r} A_i\right) \cap A_{r+1}\right] \qquad \ldots[\text{Using } (*)]$$

$$= P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) - P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \qquad \text{(Distributive Law)}$$

$$= \sum_{i=1}^{r} P(A_i) - \sum\sum_{1 \le i < j \le r} P(A_i \cap A_j) + \ldots$$
$$\ldots + (-1)^{r-1} P(A_1 \cap A_2 \cap \ldots \cap A_r) + P(A_{r+1})$$
$$- P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \qquad \ldots[\text{From } (**)]$$

$$= \sum_{i=1}^{r+1} P(A_i) - \sum\sum_{1 \le i < j \le r} P(A_i \cap A_j) + \ldots$$
$$+ (-1)^{r-1} P(A_1 \cap A_2 \cap \ldots \cap A_r)$$

$$- [\sum_{i=1}^{r} P(A_i \cap A_{r+1}) - \sum\sum_{1 \le i < j \le r} P(A_i \cap A_j \cap A_{r+1})$$
$$+ ... + (-1)^{r-1} P(A_1 \cap A_2 \cap ... \cap A_r \cap A_{r+1})] \qquad ...[\text{From } (**)]$$

$$\Rightarrow \quad P(\bigcup_{i=1}^{r+1} A_i) = \sum_{i=1}^{r+1} P(A_i) - [\sum\sum_{1 \le i < j \le r} P(A_i \cap A_j) + \sum_{i=1}^{r} P(A_i \cap A_{r+1})]$$
$$+ ... + (-1)^r P[(A_1 \cap A_2 \cap ... \cap A_{r+1})]$$
$$= \sum_{i=1}^{r+1} P(A_i) - \sum\sum_{1 \le i < j \le (r+1)} P(A_i \cap A_j)$$
$$+ ... + (-1)^r P(A_1 \cap A_2 \cap .... \cap A_{r+1})$$

Hence if (4·6) is true for $n=r$, it is also true for $n = (r+1)$. But we have proved in (*) that (4·6) is true for $n=2$. Hence by the principle of mathematical induction, it follows that (4·6) is true for all positive integral values of n.

Remarks. 1. If we write

$$P(A_i) = p_i, P(A_i \cap A_j) = p_{ij}, P(A_i \cap A_j \cap A_k) = p_{ijk}$$

and so on and

$$S_1 = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} P(A_i)$$
$$S_2 = \sum\sum_{1 \le i < j \le n} p_{ij} = \sum\sum_{1 \le i < j \le n} P(A_i \cap A_j)$$
$$S_3 = \sum\sum\sum_{1 \le i < j < k \le n} p_{ijk} \quad \text{and so on,}$$

then

$$P(\bigcup_{i=1}^{n} A_i) = S_1 - S_2 + S_3 - ... + (-1)^{n-1} S_n \qquad ...(4·6a)$$

2. If all the events A_i, $(i = 1, 2, ..., n)$ are mutually disjoint then (4·6) gives

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

3. From practical point of view the theorem can be restated in a slightly different form. Let us suppose that an event A can materialise in several mutually exclusive forms, *viz.*, $A_1, A_2, ..., A_n$ which may be regarded as that many mutually exclusive events. If A happens then any one of the events A_i, $(i = 1, 2, ..., n)$ must happen and conversely if any one of the events A_i, $(i = 1, 2, ..., n)$ happens, then A happens. Hence the probability of happening of A is the same as the probability of happening of any one of its (unspecified) mutually exclusive forms. From this point of view, the total probability theorem can be restated as follows:

The probability of happening of an event A is the sum of the probabilities of happening of its mutually exclusive forms $A_1, A_2, ..., A_n$. Symbolically,

$$P(A) = P(A_1) + P(A_2) + ... + P(A_n) \qquad (4·6b)$$

The probabilities $P(A_1), P(A_2), ..., P(A_n)$ of the mutually exclusive forms of A are known as the *partial probabilities*. Since $P(A)$ is their sum, it may be called the *total probability* of A. Hence the name of the theorem.

Theorem 4·7. *(Boole's inequality). For n events $A_1, A_2, ..., A_n$, we have*

$$(a) \quad P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1) \qquad ...(4·7)$$

$$(b) \quad P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) \qquad ...(4·7a)$$

[Delhi Univ. B.Sc. (Stat Hons.), 1992, 1989]

Proof. *(a)* $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq 1$

$$\Rightarrow \quad P(A_1 \cap A_2) \geq P(A_1) + P(A_2) - 1 \qquad (*)$$

Hence (4·7) is true for $n = 2$.

Let us now suppose that (4·7) is true for $n=r$ (say), such that

$$P\left(\bigcap_{i=1}^{r} A_i\right) \geq \sum_{i=1}^{r} P(A_i) - (r-1) \qquad (**)$$

Then

$$P\left(\bigcap_{i=1}^{r+1} A_i\right) = P\left(\bigcap_{i=1}^{r} A_i \cap A_{r+1}\right)$$

$$\geq P\left(\bigcap_{i=1}^{r} A_i\right) + P(A_{r+1}) - 1 \qquad \text{[From (*)]}$$

$$\geq \sum_{i=1}^{r} P(A_i) - (r-1) + P(A_{r+1}) - 1 \qquad \text{[From (**)]}$$

$$\Rightarrow \quad P\left(\bigcap_{i=1}^{r+1} A_i\right) \geq \sum_{i=1}^{r+1} P(A_i) - r$$

$\Rightarrow$ (4·7) is true for $n = r + 1$ also.

The result now follows by the principle of mathematical induction.

(b) Applying the inequality (4·7) to the events $\overline{A}_1, \overline{A}_2, ..., \overline{A}_n$, we get

$$P(\overline{A}_1 \cap \overline{A}_2 \cap ... \cap \overline{A}_n) \geq [P(\overline{A}_1) + P(\overline{A}_2) + ... + P(\overline{A}_n)] - (n-1)$$
$$= [1 - P(A_1)] + [1 - P(A_2)] + ... + [1 - P(A_n)] - (n-1)$$
$$= 1 - P(A_1) - P(A_2) - ... - P(A_n)$$

$$\Rightarrow \quad P(A_1) + P(A_2) + .. + P(A_n) \geq 1 - P(\overline{A}_1 \cap \overline{A}_2 \cap ... \cap \overline{A}_n).$$
$$= 1 - P(\overline{A_1 \cup A_2 \cup ... \cup A_n}).$$
$$= P(A_1 \cup A_2 \cup ... \cup A_n)$$

$$\Rightarrow \quad P(A_1 \cup A_2 \cup ... \cup A_n) \leq P(A_1) + P(A_2) + ... + P(A_n)$$

as desired.

Aliter for (b) i.e., (4·7a). We have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

$$\leq P(A_1) + P(A_2) \qquad [\because P(A_1 \cap A_2) \geq 0] \qquad ...(***)$$

Hence (4·7a) is true for $n = 2$.

Let us now suppose that (4·7a) is true for $n=r$, (say), so that

$$P\left(\bigcup_{i=1}^{r} A_i\right) \leq \sum_{i=1}^{r} P(A_i) \qquad ...(****)$$

Now

$$P\left(\bigcup_{i=1}^{r+1} A_i\right) = P\left(\bigcup_{i=1}^{r} A_i \cup A_{r+1}\right)$$

$$\leq P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) \qquad \text{[Using (***)]}$$

$$\leq \sum_{i=1}^{r} P(A_i) + P(A_{r+1}) \qquad \text{[Using (****)]}$$

$$\Rightarrow \qquad P\left(\bigcup_{i=1}^{r+1} A_i\right) \leq \sum_{i=1}^{r+1} P(A_i)$$

Hence if (4·7a) is true for $n=r$, then it is also true for $n=r+1$. But we have proved in (***) that (4·7a) is true for $n=2$. Hence by mathematical induction we conclude that (4·7a) is true for all positive integral values of n.

Theorem 4·8. *For n events* $A_1, A_2, ..., A_n$,

$$P\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j)$$

[Delhi Univ. B.Sc. (Stat Hons.), 1986]

Proof. We shall prove this theorem by the method of induction.

We know that

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$
$$- [P(A_1 \cap A_2) + P(A_2 \cap A_3) + P(A_3 \cap A_1)] + P(A_1 \cap A_2 \cap A_3)$$

$$\Rightarrow P\left(\bigcup_{i=1}^{3} A_i\right) \geq \sum_{i=1}^{3} P(A_i) - \sum\sum_{1 \leq i < j \leq 3} P(A_i \cap A_j)$$

Thus the result is true for $n=3$. Let us now suppose that the result is true for $n=r$ (say), so that

$$P\left(\bigcup_{i=1}^{r} A_i\right) \geq \sum_{i=1} P(A_i) - \sum\sum_{1 \leq i < j \leq r} P(A_i \cap A_j) \qquad ...(*)$$

Now

$$P\left(\bigcup_{i=1}^{r+1} A_i\right) = P\left(\bigcup_{i=1}^{r} A_i \cup A_{r+1}\right)$$

$$= P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) - P\left[\left(\bigcup_{i=1}^{r} A_i\right) \cap A_{r+1}\right]$$

$$= P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) - P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right]$$

$$\geq \left[\sum_{i=1}^{r} P(A_i) - \sum\sum_{1 \leq i < j < r} P(A_i \cap A_j)\right]$$

$$+ P(A_{r+1}) - P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \qquad \ldots(**)$$

[From (*)]

From Boole's inequality (*c.f.* Theorem 4·7 page 4·33), we get

$$P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \leq \sum_{i=1}^{r} P(A_i \cap A_{r+1})$$

$$\Rightarrow \quad -P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \geq -\sum_{i=1}^{r} P(A_i \cap A_{r+1})$$

$\therefore$ From (**), we get

$$P\left(\bigcup_{i=1}^{r+1} A_i\right) \geq \sum_{i=1}^{r+1} P(A_i) - \sum\sum_{1 \leq i < j \leq r} P(A_i \cap A_j) - \sum_{i=1}^{r} P(A_i \cap A_{r+1})$$

$$\Rightarrow \quad P\left(\bigcup_{i=1}^{r+1} A_i\right) \geq \sum_{i=1}^{r+1} P(A_i) - \sum\sum_{1 \leq i < j \leq r+1} P(A_i \cap A_j)$$

Hence, if the theorem is true for $n = r$, it is also true for $n = r + 1$. But we have seen that the result is true for $n = 3$. Hence by mathematical induction, the result is true for all positive integral values of n.

4·7. Multiplication Law of Probability and Conditional Probability

Theorem 4·8. *For two events A and B*

$$\left.\begin{aligned} P(A \cap B) &= P(A) \,.\, P(B \mid A), \; P(A) > 0 \\ &= P(B) \,.\, P(A \mid B), \; P(B) > 0 \end{aligned}\right\} \qquad \ldots(4\cdot8)$$

where $P(B \mid A)$ represents the conditional probability of occurrence of B when the event A has already happened and $P(A \mid B)$ is the conditional probability of happening of A, given that B has already happened.

Proof.

$$P(A) = \frac{n(A)}{n(S)} \;;\; P(B) = \frac{n(B)}{n(S)} \text{ and } P(A \cap B) = \frac{n(A \cap B)}{n(S)} \qquad (*)$$

For the conditional event $A \mid B$, the favourable outcomes must be one of the sample points of B, *i.e.*, for the event $A \mid B$, the sample space is B and out of the $n(B)$ sample points, $n(A \cap B)$ pertain to the occurrence of the event A. Hence

$$P(A \mid B) = \frac{n(A \cap B)}{n(B)}$$

Rewriting (*), we get

$$P(A \cap B) = \frac{n(B)}{n(S)} \cdot \frac{n(A \cap B)}{n(B)} = P(B) \,.\, P(A \mid B)$$

Similarly we can prove :

$$P(A \cap B) = \frac{n(A)}{n(S)} \cdot \frac{n(A \cap B)}{n(A)} = P(A) \cdot P(B \mid A)$$

Remarks. 1. $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$ and $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

Thus the conditional probabilities $P(B \mid A)$ and $P(A \mid B)$ are defined if and only if $P(A) \neq 0$ and $P(B) \neq 0$, respectively.

2. *(i)* For $P(B) > 0$, $P(A \mid B) \leq P(A)$

(ii) The conditional probability $P(A \mid B)$ is not defined if $P(B) = 0$.

(iii) $P(B \mid B) = 1$.

3. **Multiplication Law of Probability for Independent Events.** If A and B are independent then

$$P(A \mid B) = P(A) \quad \text{and} \quad P(B \mid A) = P(B)$$

Hence (4·8) gives :

$$P(A \cap B) = P(A)\, P(B) \qquad \text{...(4·8}a\text{)}$$

provided A and B are independent.

4·7·1. Extension of Multiplication Law of Probability. *For n events $A_1, A_2, \ldots, A_n$, we have*

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \ldots \times P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \qquad \text{...(4·8}b\text{)}$$

where $P(A_i \mid A_j \cap A_k \cap \ldots \cap A_l)$ represents the conditional probability of the event A_i given that the events $A_j, A_k, \ldots, A_l$ have already happened.

Proof. We have for three events A_1, A_2, and A_3

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P[A_1 \cap (A_2 \cap A_3)] \\ &= P(A_1)\, P(A_2 \cap A_3 \mid A_1) \\ &= P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \end{aligned}$$

Thus we find that (4·8b) is true for $n=2$ and $n=3$. Let us suppose that (4·8b) is true for $n=k$, so that

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \ldots P(A_k \mid A_1 \cap A_2 \cap \ldots \cap A_{k-1})$$

Now

$$\begin{aligned} P[(A_1 \cap A_2 \cap \ldots \cap A_k) \cap A_{k+1}] &= P(A_1 \cap A_2 \cap \ldots \cap A_k) \times P(A_{k+1} \mid A_1 \cap A_2 \cap \ldots \cap A_k) \\ &= P(A_1)\, P(A_2 \mid A_1) \ldots P(A_k \mid A_1 \cap A_2 \cap \ldots \cap A_{k-1}) \\ &\qquad \times P(A_{k+1} \mid A_1 \cap A_2 \cap \ldots \cap A_k) \end{aligned}$$

Thus (4·8b) is true for $n=k+1$ also. Since (4·8b) is true for $n=2$ and $n=3$, by the principle of mathematical induction, it follows that (4·8b) is true for all positive integral values of n.

Remark. If $A_1, A_2, \ldots, A_n$ are independent events then

$$P(A_2 \mid A_1) = P(A_2),\ P(A_3 \mid A_1 \cap A_2) = P(A_3) \ldots P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1}) = P(A_n)$$

Hence (4·8*b*) gives :

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2)\dots \cap P(A_n), \quad \dots(4\cdot 8c)$$

provided $A_1, A_2, \dots, A_n$ are independent.

Remark. Mutually Exclusive (Disjoint) Events and Independent Events. Let A and B be mutually exclusive (disjoint) events with positive probabilities ($P(A) > 0, P(B) > 0$), *i.e.*, both A and B are possible events such that

$$A \cap B = \phi \quad \Rightarrow \quad P(A \cap B) = P(\phi) = 0 \quad \dots(i)$$

Further, by compound probability theorem we have

$$P(A \cap B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B) \quad \dots(ii)$$

Since $P(A) \neq 0$; $P(B) \neq 0$, from *(i)* and *(ii)* we get

$$P(A \mid B) = 0 \neq P(A), \quad P(B \mid A) = 0 \neq P(B) \quad \dots(iii)$$

$\Rightarrow$ A and B are dependent events.

Hence two possible mutually disjoint events are always dependent (not independent) events.

However, if A and B are independent events with $P(A) > 0$ and $P(B) > 0$, then

$$P(A \cap B) = P(A)\ P(B) \neq 0$$

$\Rightarrow$ A and B cannot be mutually exclusive.

Hence two independent events (both of which are possible events), cannot be mutually disjoint.

4·7·2. Given n independent events A_i, (i =1,2,...,n) with respective probabilities of occurrence p_i, to find the probability of occurrence of at least one of them.

We have

$$P(A_i) = p_i \quad \Rightarrow \quad P(\bar{A}_i) = 1 - p_i\ ;\ i = 1, 2, \dots, n$$

$$[\because \overline{(A_1 \cup A_2 \cup \dots \cup A_n)} = (\bar{A}_1 \cap \bar{A}_2 \cap \dots \cap \bar{A}_n) \quad (\textit{De-Morgan's Law})]$$

Hence the probability of happening of at least one of the events is given by

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = 1 - P\,\overline{(A_1 \cup A_2 \cup \dots \cup A_n)} \quad \dots(*)$$

$$= 1 - P(\bar{A}_1 \cap \bar{A}_2 \cap \dots \cap \bar{A}_n)$$

$$= 1 - P(\bar{A}_1)\,P(\bar{A}_2) \dots P(\bar{A}_n) \quad \dots(**)$$

[*c.f.* Theorem 4·14 page 4·41]

$$= 1 - [(1 - p_1)(1 - p_2) \dots (1 - p_n)]$$

$$= \left[\sum_{i=1}^{n} p_i - \sum\sum_{\substack{i,j=1 \\ i<j}}^{n} (p_i p_j) + \sum\sum\sum_{\substack{i,j,k=1 \\ i<j<k}}^{n} (p_i p_j p_k) \right.$$

$$\left. \dots + (-1)^{n-1} (p_1 p_2 \dots p_n)\right]$$

Remark. The results in (*) and (**) are very important and are used quite often in numerical problems. Result (*) stated in words gives:

P [happening of at least one of the events $A_1, A_2, \dots, A_n$]

$= 1 - P$ (none of the events $A_1, A_2, \dots, A_n$ happens) or equivalently,

P {none of the given events happens}

$= 1 - P$ {at least one of them happens}.

Theorem 4·9. *For any three events A, B and C*

$$P(A \cup B \mid C) = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$$

Proof. We have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\Rightarrow P[(A \cap C) \cup (B \cap C)] = P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)$$

Dividing both sides by $P(C)$, we get

$$\frac{P[(A \cap C) \cup (B \cap C)]}{P(C)} = \frac{P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)}{P(C)}, P(C) > 0$$

$$= \frac{P(A \cap C)}{P(C)} + \frac{P(B \cap C)}{P(C)} - \frac{P(A \cap B \cap C)}{P(C)}$$

$$\Rightarrow \quad \frac{P[(A \cup B) \cap C]}{P(C)} = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$$

$$\Rightarrow \quad P[(A \cup B) \mid C] = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$$

Theorem 4·10. *For any three events A, B and C*

$$P(A \cap \bar{B} \mid C) + P(A \cap B \mid C) = P(A \mid C)$$

Proof. $P(A \cap \bar{B} \mid C) + P(A \cap B \mid C)$

$$= \frac{P(A \cap \bar{B} \cap C)}{P(C)} + \frac{P(A \cap B \cap C)}{P(C)}$$

$$= \frac{P(A \cap \bar{B} \cap C) + P(A \cap B \cap C)}{P(C)}$$

$$= \frac{P(A \cap C)}{P(C)} = P(A \mid C)$$

Theorem 4·11. *For a fixed B with $P(B) > 0$, $P(A \mid B)$ is a probability function.* **[Delhi Univ. B.Sc. (Stat. Hons.), 1991; (Maths Hons.), 1992]**

Proof.

(i) $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} \geq 0$

(ii) $P(S \mid B) = \dfrac{P(S \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1$

(iii) If $\{A_n\}$ is any finite or infinite sequences of disjoint events, then

$$P[\bigcup_n A_n \mid B] = \frac{P[(\bigcup_n A_n) \cap B]}{P(B)} = \frac{P[(\bigcup_n A_n . B)]}{P(B)}$$

$$= \frac{\sum_n P(A_n B)}{P(B)} = \sum_n \left[\frac{P(A_n B)}{P(B)}\right] = \sum_n P(A_n \mid B)$$

Hence the theorem.

Remark. For given B satisfying $P(B) > 0$, the conditional probability $P[\cdot|B]$ also enjoys the same properties as the unconditional probability.

For example, in the usual notations, we have:

(*i*) $P\,[\phi \mid B] = 0$

(*ii*) $P\,[\,\overline{A} \mid B] = 1 - P\,[A \mid B]$

(*iii*) $P[\bigcup_{i=1}^{n} A_i \mid B] = \sum_{i=1}^{n} P[A_i \mid B]$,

where $A_1, A_2, ..., A_n$ are mutually disjoint events.

(*iv*) $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P\,(A_2 \mid B) - P(A_1\, A_2 \mid B)$

(*v*) If $E \subset F$, then $P(E \mid B) \leq P(F \mid B)$

and so on.

The proofs of results (*iv*) and (*v*) are given in theorems 4·9 and 4·13 respectively. Others are left as exercises to the reader.

Theorem 4·12. *For any three events, A, B and C defined on the sample space S such that $B \subset C$ and $P(A) > 0$,*

$$P(B \mid A) \leq P(C \mid A)$$

Proof.
$$P(C \mid A) = \frac{P(C \cap A)}{P(A)} \qquad \text{(By definition)}$$

$$= \frac{P[B \cap C \cap A) \cup (\overline{B} \cap C \cap A)}{P(A)}$$

$$= \frac{P[B \cap C \cap A)}{P(A)} + \frac{P(\overline{B} \cap C \cap A)}{P(A)} \qquad \text{(Using axiom 3)}$$

$$= P[(B \cap C|A) + (\overline{B} \cap C \cap A)]$$

Now
$$B \subset C \quad \Rightarrow \quad B \cap C = B$$

$\therefore$
$$P(C \mid A) = P(B \mid A) + P(\overline{B} \cap C|A)$$

$\Rightarrow$
$$P(C \mid A) \geq P(B \mid A)$$

4·7·3. Independent Events. *An event B is said to be independent (or statistically independent) of event A, if the conditional probability of B given A i.e., P (B | A) is equal to the unconditional probability of B, i.e., if*

$$P\,(B \mid A) = P\,(B)$$

Since

$$P\,(A \cap B) = P\,(B \mid A)\,P\,(A) = P\,(A \mid B)\,P\,(B)$$

and since $P\,(B \mid A) = P\,(B)$ when B is independent of A, we must have $P\,(A \mid B) = P\,(A)$ or it follows that A is also independent of B. Hence the events A and B are independent if and only if

$$P\,(A \cap B) = P\,(A)\,P\,(B) \qquad ...(4\cdot9)$$

4·7·4. Pairwise Independent Events

Definition. *A set of events $A_1, A_2, ..., A_n$ are said to be pair-wise independent if*

$$P\,(A_i \cap A_j) = P\,(A_i)\,P\,(A_j) \;\; \forall \;\; i \neq j \qquad ...(4\cdot10)$$

4·7·5. Conditions for Mutual Independence of n Events. Let S denote the sample space for a number of events. The events in S are said to be mutually indepdendent if the probability of the simultaneous occurrence of (any) finite number of them is equal to the product of their separate probabilities.

If $A_1, A_2, ..., A_n$ are n events, then for their mutual independence, we should have

(i) $P(A_i \cap A_j) = P(A_i) P(A_j), \ (i \neq j \ ; i, j = 1, 2, ..., n)$

(ii) $P(A_i \cap A_j \cap A_k) = P(A_i) P(A_j) P(A_k), (i \neq j \neq k \ ; i \ ; j \ ; k = 1, 2, ..., n)$

$$\vdots \qquad\qquad \vdots$$

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) P(A_2) ... P(A_n)$$

It is interesting to note that the above equations give respectively ${}^nC_2, {}^nC_3, ..., {}^nC_n$ conditions to be satisfied by $A_1, A_2, ..., A_n$.

Hence the total number of conditions for the mutual independence of $A_1, A_2, ..., A_n$ is ${}^nC_2 + {}^nC_3 + ... + {}^nC_n$.

Since ${}^nC_0 + {}^nC_1 + {}^nC_2 + ... + {}^nC_n = 2^n$, we get the required number of conditions as $(2^n - 1 - n)$.

In particular for three events A_1, A_2 and A_3, $(n = 3)$, we have the following $2^3 - 1 - 3 = 4$, conditions for their mutual independence.

$$\begin{aligned}
P(A_1 \cap A_2) &= P(A_1) \ P(A_2) \\
P(A_2 \cap A_3) &= P(A_2) \ P(A_3) \\
P(A_1 \cap A_3) &= P(A_1) \ P(A_3) \\
P(A_1 \cap A_2 \cap A_3) &= P(A_1) \ P(A_2) \ P(A_3)
\end{aligned} \qquad ...(4{\cdot}11)$$

Remarks. 1. It may be observed that pairwise or mutual independence of events $A_1, A_2, ..., A_n$, is defined only when $P(A_i) \neq 0$, for $i = 1, 2, ..., n$.

2. If the events A and B are such that $P(A_i) \neq 0$, $P(B) \neq 0$ and A is independent of B, then B is independent of A.

Proof. We are given that

$$P(A \mid B) = P(A)$$

$$\Rightarrow \quad \frac{P(A \cap B)}{P(B)} = P(A)$$

$$\Rightarrow \quad P(A \cap B) = P(A) \ P(B)$$

$$\Rightarrow \quad \frac{P(B \cap A)}{P(A)} = P(B) \qquad [\because P(A) \neq 0 \text{ and } A \cap B = B \cap A]$$

$$\Rightarrow \quad P(B \mid A) = P(B),$$

which by definition of independent events, means that B is independent of A.

3. It may be noted that pairwise independence of events does not imply their mutual independence. For illustrations, see Examples 4·50 and 4·51.

Theorem 4·13. *If A and B are independent events then A and $\overline{B}$ are also independent events.*

Proof. By theorem 4·4, we have

$$P(A \cap \bar{B}) = P(A) - P(A \cap B)$$
$$= P(A) - P(A)P(B) \quad [\because A \text{ and } B \text{ are independent}]$$
$$= P(A)[1 - P(B)]$$
$$= P(A)\,P(\bar{B})$$

$\Rightarrow$ A and $\bar{B}$ are independent events.

Aliter. $P(A \cap B) = P(A)\,P(B) = P(A)\,P(B \mid A) = P(B)\,P(A \mid B)$

i.e., $P(B \mid A) = P(B) \Rightarrow B$ is independent of A.

also $P(A \mid B) = P(A) \Rightarrow A$ is independent of B.

Also $P(B \mid A) + P(\bar{B} \mid A) = 1 \Rightarrow P(B) + P(\bar{B} \mid A) = 1$

or $P(\bar{B} \mid A) = 1 - P(B) = P(\bar{B})$

$\therefore$ $\bar{B}$ is independent of A and by symmetry we say that A is independent of $\bar{B}$. Thus A and $\bar{B}$ are independent events.

Remark. Similarly, we can prove that if A and B are independent events then $\bar{A}$ and B are also independent events.

Theorem 4·14. *If A and B are independent events then $\bar{A}$ and $\bar{B}$ are also independent events.*

Proof. We are given $P(A \cap B) = P(A)\,P(B)$

Now $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$

$$= 1 - [P(A) + P(B) - P(A \cap B)]$$
$$= 1 - [P(A) + P(B) - P(A)P(B)]$$
$$= 1 - P(A) - P(B) + P(A)P(B)$$
$$= [1 - P(B)] - P(A)[1 - P(B)]$$
$$= [1 - P(A)][1 - P(B)] = P(\bar{A})P(\bar{B})$$

$\therefore$ $\bar{A}$ and $\bar{B}$ are independent events.

Aliter. We know

$$P(\bar{A} \mid \bar{B}) + P(A \mid \bar{B}) = 1$$

$\Rightarrow$ $P(\bar{A} \mid \bar{B}) + P(A) = 1$ (*c.f.* Theorem 4·13)

$\Rightarrow$ $P(\bar{A} \mid \bar{B}) = 1 - P(A) = P(\bar{A})$

$\therefore$ $\bar{A}$ and $\bar{B}$ are independent events.

Theorem 4·15. *If A, B, C are mutually independent events then $A \cup B$ and C are also independent.*

Proof. We are required to prove:

$$P[(A \cup B) \cap C] = P(A \cup B)P(C)$$

L.H.S. $= P[(A \cap C) \cup (B \cap C)]$ [Distributive Law]

$$= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)$$
$$= P(A)\,P(C) + P(B)\,P(C) - P(A)\,P(B)\,P(C)$$

[$\because$ A, B and C are mutually independent]

$$= P(C)\,[P(A) + P(B) - P(A \cap B)]$$

$= P(C)\,P(A \cup B) =$ R.H.S.

Hence $(A \cup B)$ and C are independent.

Theorem 4·16. *If A, B and C are random events in a sample space and if A, B and C are pairwise independent and A is independent of $(B \cup C)$, then A, B and C are mutually independent.*

Proof. We are given

$$\left.\begin{aligned} P(A \cap B) &= P(A)P(B) \\ P(B \cap C) &= P(B)P(C) \\ P(A \cap C) &= P(A)P(C) \\ P[A \cap (B \cup C)] &= P(A)P(B \cup C) \end{aligned}\right\} \quad \ldots (*)$$

Now $\quad P[A \cap (B \cup C)] = P[(A \cap B) \cup (A \cap C)]$

$$= P(A \cap B) + P(A \cap C) - P[(A \cap B) \cap (A \cap C)]$$

$$= P(A).P(B) + P(A).P(C) - P(A \cap B \cap C) \quad \ldots (**)$$

and $P(A)\,P(B \cup C) = P(A)\,[P(B) + P(C) - P(B \cap C)]$

$$= P(A).P(B) + P(A)P(C) - P(A)P(B \cap C) \quad \ldots (***)$$

From (**) and (***), on using (*), we get

$P(A \cap B \cap C) = P(A)P(B \cap C) = P(A)P(B)P(C)$

Hence A, B, C are mutually independent.

Theorem 4·17. *For any two events A and B,*

$$P(A \cap B) \le P(A) \le P(A \cup B) \le P(A) + P(B)$$

[Patna Univ. B.A.(Stat. Hons.), 1992; Delhi Univ. B.Sc.(Stat. Hons.), 1989]

Proof. We have

$$A = (A \cap \bar{B}) \cup (A \cap B)$$

Using axiom 3, we have

$$P(A) = P[(A \cap \bar{B}) \cup (A \cap B)] = P(A \cap \bar{B}) + P(A \cap B)$$

Now $\quad P[(A \cap \bar{B}) \ge 0$ (From axiom 1)

$\therefore \quad P(A) \ge P(A \cap B) \quad \ldots (*)$

Similarly $\quad P(B) \ge P(A \cap B)$

$\Rightarrow \quad P(B) - P(A \cap B) \ge 0$

Now $\quad P(A \cup B) = P(A) + [P(B) - P(A \cap B)] \quad \ldots (**)$

$\therefore \quad P(A \cup B) \ge P(A) \quad \Rightarrow \quad P(A) \le P(A \cup B) \quad \ldots (***)$

Also $\quad P(A \cup B) \le P(A) + P(B)$ [From (**)]

Hence from (*), (**) and (***), we get

$$P(A \cap B) \le P(A) \le P(A \cup B) \le P(A) + P(B)$$

Aliter. Since $A \cap B \subset A$, by Theorem 4·6 *(ii)* page 4·30, we get

$$P(A \cap B) \le P(A).$$

Also $A \subset (A \cup B) \quad \Rightarrow \quad P(A) \le P(A \cup B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\le P(A) + P(B) \qquad [\because P(A \cap B) \ge 0]$$

Combining the above results, we get

$$P(A \cap B) \le P(A) \le P(A \cup B) \le P(A) + P(B)$$

Example 4·12. *Two dice, one green and the other red, are thrown. Let A be the event that the sum of the points on the faces shown is odd, and B be the event of at least one ace (number '1').*

(a) Describe the (i) complete sample space, (ii) events A, B, $\overline{B}$, $A \cap B$, $A \cup B$, and $A \cap \overline{B}$ and find their probabilities assuming that all the 36 sample points have equal probabilities.

(b) Find the probabilities of the events :

(i) $(\overline{A} \cup \overline{B})$ *(ii)* $(\overline{A} \cap \overline{B})$ *(iii)* $(A \cap \overline{B})$ *(iv)* $(\overline{A} \cap B)$ *(v)* $(\overline{A \cap B})$ *(vi)* $(\overline{A} \cup B)$ *(vii)* $(\overline{A \cup B})$ *(viii)* $\overline{A} \cap (A \cup B)$ *(ix)* $A \cup (\overline{A} \cap B)$ *(x)* $(A \mid B)$ and $(B \mid A)$, and *(xi)* $(\overline{A} \mid \overline{B})$ *and* $(\overline{B} \mid \overline{A})$ . .

Solution., *(a)* The sample space consists of the 36 elementary events .

(1,1) ; (1,2) ; (1,3) ; (1,4) ; (1,5) ; (1,6)
(2,1) ; (2,2) ; (2,3) ; (2,4) ; (2,5) ; (2,6)
(3,1) ; (3,2) ; (3,3) ; (3,4) ; (3,5) ; (3,6)
(4,1) ; (4,2) ; (4,3) ; (4,4) ; (4,5) ; (4,6)
(5,1) ; (5,2) ; (5,3) ; (5,4) ; (5,5) ; (5,6)
(6,1) ; (6,2) ; (6,3) ; (6,4) ; (6,5) ; (6,6)

where, for example, the ordered pair (4, 5) refers to the elementary event that the green die shows 4 and and the red die shows 5.

A = The event that the sum of the numbers shown by the two dice is odd.
= {(1,2); (2,1); (1,4); (2,3); (3,2); (4,1); (1,6); (2,5)
(3,4); (4,3); (5,2); (6,1); (3,6); (4,5); (5,4); (6,3)
(5,6); (6,5)} and therefore

$$P(A) = \frac{n(A)}{n(S)} = \frac{18}{36}$$

B = The event that at least one face is 1,
= {(1,1) ; (1,2) ; (1,3) ; (1,4) ; (1,5) ; (1,6)
(2,1) ; (3,1) ; (4,1) ; (5,1) ; (6,1)} and therefore

$$P(B) = \frac{n(B)}{n(S)} = \frac{11}{36}$$

$\overline{B}$ = The event that each of the face obtained is not an ace.
= {(2,2); (2,3); (2,4); (2,5); (2,6); (3,2); (3,3);
(3,4) ; (3,5) ; (3,6); (4,2) ; (4,3); (4,4) ; (4,5);
(4,6) ; (5,2) ; (5,3); (5,4) ; (5,5); (5,6); (6,2) ;
(6,3) ; (6,4) ; (6,5); (6,6)} and therefore

$$P(\overline{B}) = \frac{n(\overline{B})}{n(S)} = \frac{25}{36}$$

$A \cap B$ = The event that sum is odd and at least one face is an ace.
= {(1,2); (2,1); (1,4); (4,1); (1,6); (6,1)}

$$\therefore \quad P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

$$A \cup B = \{(1,2);(2,1);(1,4);(2,3);(3,2);(4,1);(1,6);(2,5)\\ (3,4);(4,3);(5,2);(6,1);(3,6);(4,5);(5,4);(6,3)\\ (5,6);(6,5);(1,1);(1,3);(1,5);(3,1);(5,1)\}$$

$$\therefore \quad P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{23}{36}$$

$$A \cap \bar{B} = \{(2,3);(3,2);(2,5);(3,4);(3,6);(4,3);(4,5);(5,2)\\ (5,4);(5,6);(6,3);(6,5)\}$$

$$P(A \cap \bar{B}) = \frac{n(A \cap \bar{B})}{n(S)} = \frac{12}{36} = \frac{1}{3}$$

(b) (i) $P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - \frac{1}{6} = \frac{5}{6}$

(ii) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - \frac{23}{36} = \frac{13}{36}$

(iii) $P(A \cap \bar{B}) = P(A) - P(A \cap B) = \frac{18}{36} - \frac{6}{36} = \frac{12}{36} = \frac{1}{3}$

(iv) $P(\bar{A} \cap B) = P(B) - P(A \cap B) = \frac{11}{36} - \frac{6}{36} = \frac{5}{36}$

(v) $P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - \frac{1}{6} = \frac{5}{6}$

(vi) $P(\bar{A} \cup B) = P(\bar{A}) + P(B) - P(\bar{A} \cap B)$

$$= \left(1 - \frac{18}{36}\right) + \frac{11}{36} - \frac{5}{36} = \frac{2}{3}$$

(vii) $P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - \frac{23}{36} = \frac{13}{36}$

(viii) $P[\bar{A} \cap (A \cup B)] = P[(A \cap \bar{A}) \cup (\bar{A} \cap B)]$

$$= P(\bar{A} \cap B) = \frac{5}{36} \qquad [\because A \cap \bar{A} = \phi]$$

(ix) $P[A \cup (\bar{A} \cap B)] = P(A) + P(\bar{A} \cap B) - P(A \cap \bar{A} \cap B)$

$$= P(A) + P(\bar{A} \cap B) = \frac{18}{36} + \frac{5}{36} = \frac{23}{36}$$

(x) $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{6/36}{11/36} = \frac{6}{11}$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{6/36}{18/36} = \frac{6}{18} = \frac{1}{3}$$

(xi) $P(\bar{A} \mid \bar{B}) = \frac{P(\bar{A} \cap \bar{B})}{P(\bar{B})} = \frac{13/36}{25/36} = \frac{13}{25}$

$$P(\bar{B} \mid \bar{A}) = \frac{P(\bar{A} \cap \bar{B})}{P(\bar{A})} = \frac{13/36}{18/36} = \frac{13}{18}$$

Example 4·13. *If two dice are thrown, what is the probability that the sum is (a) greater than 8, and (b) neither 7 nor 11?*

Solution. *(a)* If S denotes the sum on the two dice, then we want $P(S > 8)$. The required event can happen in the following mutually exclusive ways:
(i) $S = 9$ *(ii)* $S = 10$ *(iii)* $S = 11$ *(iv)* $S = 12$.
Hence by addition theorem of probability

$$P(S > 8) = P(S = 9) + P(S = 10) + P(S = 11) + P(S = 12)$$

In a throw of two dice, the sample space contains $6^2 = 36$ points.

The number of favourable cases can be enumerated as follows:

$S = 9$: (3, 6), (6, 3), (4, 5), (5, 4), *i.e.*, 4 sample points.

$$\therefore \quad P(S = 9) = \frac{4}{36}$$

$S = 10$: (4,6), (6,4), (5,5), *i.e.*, 3 sample points.

$$\therefore \quad P(S = 10) = \frac{3}{36}$$

$S = 11$: (5, 6), (6, 5), *i.e.*, 2 sample points.

$$\therefore \quad P(S = 11) = \frac{2}{36}$$

$S = 12$: (6, 6), *i.e.*, 1 sample point.

$$\therefore \quad P(S = 12) = \frac{1}{36}$$

$$\therefore \quad P(S > 8) = \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{10}{36} = \frac{5}{18}$$

(b) Let A denote the event of getting the sum of 7 and B denote the event of getting the sum of 11 with a pair of dice.

$S = 7$: (1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3), *i.e.*, 6 distinct sample points.

$$\therefore \quad P(A) = P(S = 7) = \frac{6}{36} = \frac{1}{6}$$

$S = 11$: (5, 6), (6, 5), $P(B) = P(S = 11) = \frac{2}{36} = \frac{1}{18}$

$$\therefore \text{ Required probability} = P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B)$$

$$= 1 - [P(A) + P(B)]$$

($\because$ A and B are disjoint events)

$$= 1 - \frac{1}{6} - \frac{1}{18} = \frac{7}{9}$$

Example 4·14. *An urn contains 4 tickets numbered 1, 2, 3, 4 and another contains 6 tickets numbered 2, 4, 6, 7, 8, 9. If one of the two urns is chosen at random and a ticket is drawn at random from the chosen urn, find the probabilities that the ticket drawn bears the number (i) 2 or 4, (ii) 3, (iii) 1 or 9*

[Calicut Univ. B.Sc., 1992]

Solution. *(i)* Required event can happen in the following mutually exclusive ways:

(I) First urn is chosen and then a ticket is drawn.

(II) Second urn is chosen and then a ticket is drawn.

Since the probability of choosing any urn is ½, the required probability 'p' is given by

$$p = P(I) + P(II)$$

$$= \frac{1}{2} \times \frac{2}{4} + \frac{1}{2} \times \frac{2}{6} = \frac{5}{12}$$

(ii) Required probability $= \frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times 0 = \frac{1}{8}$

($\because$ in the 2nd urn there is no ticket with number 3)

(iii) Required probability $= \frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{6} = \frac{5}{24}$

Example 4·15. *A card is drawn from a well-shuffled pack of playing cards. What is the probability that it is either a spade or an ace ?*

Solution. The equiprobable sample space S of drawing a card from a well-shuffled pack of playing cards consists of 52 sample points.

If A and B denote the events of drawing a 'spade card' and 'an ace' respectively then A consists of 13 sample points and B consists of 4 sample points so that,

$$P(A) = \frac{13}{52} \quad \text{and} \quad P(B) = \frac{4}{52}$$

The compound event $A \cap B$ consists of only one sample point, *viz.*, ace of spade so that,

$$P(A \cap B) = \frac{1}{52}$$

The probability that the card drawn is either a spade or an ace is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}$$

Example 4·16. *A box contains 6 red, 4 white and 5 black balls. A person draws 4 balls from the box at random. Find the probability that among the balls drawn there is at least one ball of each colour.* **(Nagpur Univ. B.Sc., 1992)**

Solution. The required event E that 'in a draw of 4 balls from the box at random there is at least one ball of each colour', can materialise in the following mutually disjoint ways :

(i) 1 Red, 1 White, 2 Black balls

(ii) 2 Red, 1 White, 1 Black balls

(iii) 1 Red, 2 White, 1 Black balls.

Hence by the addition theorem of probability, the required probability is given by

$$P(E) = P(i) + P(ii) + P(iii)$$
$$= \frac{{}^6C_1 \times {}^4C_1 \times {}^5C_2}{{}^{15}C_4} + \frac{{}^6C_2 \times {}^4C_1 \times {}^5C_1}{{}^{15}C_4} + \frac{{}^6C_1 \times {}^4C_2 \times {}^5C_1}{{}^{15}C_4}$$
$$= \frac{1}{{}^{15}C_4}\left[6 \times 4 \times 10 + 15 \times 4 \times 5 + 6 \times 6 \times 5\right]$$
$$= \frac{4\,!}{15 \times 14 \times 13 \times 12}\left[240 + 300 + 180\right]$$
$$= \frac{24 \times 720}{15 \times 14 \times 13 \times 12} = 0{\cdot}5275$$

Example 4·17. *Why does it pay to bet consistently on seeing 6 at least once in 4 throws of a die, but not on seeing a double six at least once in 24 throws with two dice? (de Mere's Problem).*

Solution. The probability of getting a '6' in a throw of die $= 1/6$.

$\therefore$ The probability of not getting a '6' in a throw of die

$$= 1 - 1/6 = 5/6 .$$

By compound probability theorem, the probability that in 4 throws of a die no '6' is obtained $= (5/6)^4$

Hence the probability of obtaining '6' at least once in 4 throws of a die $= 1 - (5/6)^4 = 0{\cdot}516$

Now, if a trial consists of throwing two dice at a time, then the probability of getting a 'double' of '6' in a trial $= 1/36$.

Thus the probability of not getting a 'double of 6' in a trial $= 35/36$.

The probability that in 24 throws, with two dice each, no 'double of 6' is obtained $= (35/36)^{24}$

Hence the probability of getting a 'double of 6' at least once in 24 throws $= 1 - (35/36)^{24} = 0{\cdot}491$.

Since the probability in the first case is greater than the probability in the second case, the result follows.

Example 4·18. *A problem in Statistics is given to the three students A, B and C whose chances of solving it are* $1/2$, $3/4$, *and* $1/4$ *respectively.*

What is the probability that the problem will be solved if all of them try independently? **[Madurai Kamraj Univ. B.Sc.,1986; Delhi Univ. B.A.,1991]**

Solution. Let A, B, C denote the events that the problem is solved by the students A, B, C respectively. Then

$$P(A) = \frac{1}{2},\ P(B) = \frac{3}{4} \text{ and } P(C) = \frac{1}{4}$$

The problem will be solved if at least one of them solves the problem. Thus we have to calculate the probability of occurrence of at least one of the three events A, B, C, *i.e.*, $P(A \cup B \cup C)$.

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ &\qquad - P(B \cap C) + P(A \cap B \cap C) \\ &= P(A) + P(B) + P(C) - P(A)P(B) - P(A)P(C) \\ &\qquad - P(B)P(C) + P(A)P(B)P(C) \end{aligned}$$

$(\because A, B, C \text{ are independent events.})$

$$\begin{aligned} &= \frac{1}{2} + \frac{3}{4} + \frac{1}{4} - \frac{1}{2}\cdot\frac{3}{4} - \frac{3}{4}\cdot\frac{1}{4} \\ &\qquad - \frac{1}{2}\cdot\frac{1}{4} + \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{4} \\ &= \frac{29}{32} \end{aligned}$$

Aliter. $P(A \cup B \cup C) = 1 - P(\overline{A \cup B \cup C})$

$$= 1 - P(\bar{A} \cap \bar{B} \cap \bar{C})$$

$$= 1 - P(\bar{A})P(\bar{B})P(\bar{C})$$

$$= 1 - \left(1 - \frac{1}{2}\right)\left(1 - \frac{3}{4}\right)\left(1 - \frac{1}{4}\right)$$

$$= \frac{29}{32}$$

Example 4·19. *If $A \cap B = \phi$, then show that* ...(*)

$$P(A) \le P(\bar{B})$$

[Delhi Univ. B.Sc. (Maths Hons.) 1987]

Solution. We have

$$A = (A \cap B) \cup (A \cap \bar{B})$$

$$= \phi \cup (A \cap \bar{B}) \qquad \text{[Using *]}$$

$$= A \cap \bar{B}$$

$\Rightarrow$ $A \subseteq \bar{B}$

$\Rightarrow$ $P(A) \le P(\bar{B})$

as desired.

Aliter. Since $A \cap B = \phi$, we have $A \subset \bar{B}$, which implies that $P(A) \le P(\bar{B})$.

Example 4·20. *Let A and B be two events such that*

$$P(A) = \frac{3}{4} \text{ and } P(B) = \frac{5}{8}$$

show that

(a) $P(A \cup B) \ge \frac{3}{4}$

(b) $\frac{3}{8} \le P(A \cap B) \le \frac{5}{8}$

[Delhi Univ. B.Sc. Stat (Hons.) 1986,1988]

Solution. *(i)* We have

$$A \subset (A \cup B)$$

$\Rightarrow$ $P(A) \le P(A \cup B)$

$\Rightarrow$ $\frac{3}{4} \le P(A \cup B)$

$\Rightarrow$ $P(A \cup B) \ge \frac{3}{4}$

(ii) $A \cap B \subseteq B$

$\Rightarrow$ $P(A \cap B) \le P(B) = \frac{5}{8}$ ...*(i)*

Also $P(A \cup B) = P(A) + P(B) - P(A \cap B) \le 1$

$\Rightarrow$ $\frac{3}{4} + \frac{5}{8} - 1 \le P(A \cap B)$

$$\Rightarrow \qquad \frac{6+5-8}{8} \le P(A \cap B)$$

$$\Rightarrow \qquad \frac{3}{8} \le P(A \cap B) \qquad \qquad ...(ii)$$

From *(i)* and *(ii)* we get

$$\frac{3}{8} \le P(A \cap B) \le \frac{5}{8}$$

Example 4·21. *(Chebychev's Problem). What is the chance that two numbers, chosen at random, will be prime to each other ?*

Solution. If any number 'a' is divided by a prime number 'r', then the possible remainders are 0, 1, 2, ...r–1. Hence the chance that 'a' is divisible by r is $1/r$ (because the only case favourable to this is remainder being 0). Similarly, the probability that any number 'b' chosen at random is divisible by r is $1/r$. Since the numbers a and b are chosen at random, the probability that none of them is divisible by 'r' is given (by compound probability theorem) by :

$$\left(1-\frac{1}{r}\right) \times \left(1-\frac{1}{r}\right) = \left(1-\frac{1}{r}\right)^2; \quad r = 2, 3, 5, 7, \ldots$$

Hence the required probability that the two numbers chosen at random are prime to each other is given by

$$P = \prod_r \left(1-\frac{1}{r}\right)^2, \quad \text{where } r \text{ is a prime number.}$$

$$= \frac{6}{\pi^2} \qquad \qquad \text{(From trigonometry)}$$

Example 4·22. *A bag contains 10 gold and 8 silver coins. Two successive drawings of 4 coins are made such that : (i) coins are replaced before the second trial, (ii) the coins are not replaced before the second trial. Find the probability that the first drawing will give 4 gold and the second 4 silver coins.*

[Allahabad Univ. B.Sc., 1987]

Solution. Let A denote the event of drawing 4 gold coins in the first draw and B denote the event of drawing 4 silver coins in the second draw. Then we have to find the probability of $P(A \cap B)$.

(i) Draws with replacement. If the coins drawn in the first draw are replaced back in the bag before the second draw then the events A and B are independent and the required probability is given (using the multiplication rule of probability) by the expression

$$P(A \cap B) = P(A) \,.\, P(B) \qquad \qquad ...(*)$$

Ist draw. Four coins can be drawn out of 10+8=18 coins in ${}^{18}C_4$ ways, which gives the exhaustive number of cases. In order that all these coins are of gold, they must be drawn out of the 10 gold coins and this can be done in ${}^{10}C_4$ ways. Hence

$$P(A) = {}^{10}C_4 \,/\, {}^{18}C_4$$

2nd draw. When the coins drawn in the first draw are replaced before the 2nd draw, the bag contains 18 coins. The probability of drawing 4 silver coins in the 2nd draw is given by $P(B) = {}^{8}C_4 \ / \ {}^{18}C_4$.

Substituting in (*), we have

$$P(A \cap B) = \frac{{}^{10}C_4}{{}^{18}C_4} \times \frac{{}^{8}C_4}{{}^{18}C_4}$$

(ii) Draws without replacement. If the coins drawn are not eplaced back before the second draw, then the events A and B are not independent and the required probability is given by

$$P(A \cap B) = P(A) \, . \, P(B \mid A) \qquad \text{...(**)}$$

As discussed in part *(i)*, $P(A) = {}^{10}C_4 \ / \ {}^{18}C_4$.

Now, if the 4 gold coins which were drawn in the first draw are not replaced back, there are $18 - 4 = 14$ coins left in the bag and $P(B \mid A)$ is the probability of drawing 4 silver coins from the bag containing 14 coins out of which 6 are gold coins and 8 are silver coins.

Hence $\quad P(B \mid A) = {}^{8}C_4 \ / \ {}^{14}C_4$

Substituting in (**) we get

$$P(A \cap B) = \frac{{}^{10}C_4}{{}^{18}C_4} \times \frac{{}^{8}C_4}{{}^{14}C_4}$$

Example 4·23. *A consignment of 15 record players contains 4 defectives. The record players are selected at random, one by one, and examined. Those examined are not put back. What is the probability that the 9th one examined is the last defective?*

Solution. Let A be the event of getting exactly 3 defectives in examination of 8 record players and let B the event that the 9th piece examined is a defective one.

Since it is a problem of sampling without replacement and since there are 4 defectives out of 15 record players, we have

$$P(A) = \frac{\binom{4}{3} \times \binom{11}{5}}{\binom{15}{8}}$$

$P(B \mid A)$ = Probability that the 9th examined record player is defective given that there were 3 defectives in the first 8 pieces examined.

$= 1/7,$

since there is only one defective piece left among the remaining $15 - 8 = 7$ record players.

Hence the required probability is

$$P(A \cap B) = P(A) \, . \, P(B \mid A)$$

$$= \frac{\binom{4}{3} \times \binom{11}{5}}{\binom{15}{8}} \times \frac{1}{7} = \frac{8}{195}$$

Example 4·24. *p is the probability that a man aged x years will die in a year. Find the probability that out of n men $A_1, A_2, ..., A_n$ each aged x, A_1 will die in a year and will be the first to die.* **[Delhi Univ. B.Sc., 1985]**

Solution. Let E_i, $(i = 1, 2, ..., n)$ denote the event that A_i dies in a year. Then

$P(E_i) = p, (i = 1, 2, ..., n)$ and $P(\bar{E}_i) = 1 - p.$

The probability that none of n men $A_1, A_2, ..., A_n$ dies in a year

$$= P(\bar{E}_1 \cap \bar{E}_2 \cap ... \cap \bar{E}_n) = P(\bar{E}_1) P(\bar{E}_2) ... P(\bar{E}_n)$$

(By compound probability theorem)

$$= (1 - p)^n$$

$\therefore$ The probability that at least one of $A_1, A_2, ..., A_n$, dies in a year

$$= 1 - P(\bar{E}_1 \cap \bar{E}_2 \cap ... \cap \bar{E}_n) = 1 - (1 - p)^n$$

The probability that among n men, A_1 is the first to die is $1/n$ and since this event is independent of the event that at least one man dies in a year, required probability is

$$\frac{1}{n}\left[1 - (1 - p)^n\right]$$

Example 4·25. *The odds against Manager X settling the wage dispute with the workers are 8:6 and odds in favour of manager Y settling the same dispute are 14:16.*

(i) What is the chance that neither settles the dispute, if they both try, independently of each other?

(ii) What is the probability that the dispute will be settled?

Solution. Let A be the event that the manager X will settle the dispute and B be the event that the Manager Y will settle the dispute. Then clearly

$$P(\bar{A}) = \frac{8}{8+6} = \frac{4}{7} \quad \Rightarrow \quad P(A) = 1 - P(\bar{A}) = \frac{6}{14} = \frac{3}{7}$$

$$P(B) = \frac{14}{14+16} = \frac{7}{15} \quad \Rightarrow \quad P(\bar{B}) = 1 - P(B) = \frac{16}{14+16} = \frac{8}{15}$$

The required probability that neither settles the dispute is given by :

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B}) = \frac{4}{7} \times \frac{8}{15} = \frac{32}{105}$$

[Since A and B are independent $\Rightarrow$ $\bar{A}$ and $\bar{B}$ are also independent]

(ii) The dispute will be settled if at least one of the managers X and Y settles the dispute. Hence the required probability is given by:

$$P(A \cup B) = \text{Prob. [At least one of X and Y settles the dispute]}$$

$$= 1 - \text{Prob. [None settles the dispute]}$$

$$= 1 - P(\bar{A} \cap \bar{B}) = 1 - \frac{32}{105} = \frac{73}{105}$$

Example 4·26. *The odds that person X speaks the truth are 3:2 and the odds that person Y speaks the truth are 5:3. In what percentage of cases are they likely to contradict each other on an identical point.*

Solution. Let us define the events:

A : X speaks the truth, B : Y speaks the truth

Then $\bar{A}$ and $\bar{B}$ represent the complementary events that X and Y tell a lie respectively. We are given:

$$P(A) = \frac{3}{3+2} = \frac{3}{5} \quad \Rightarrow \quad P(\bar{A}) = 1 - \frac{3}{5} = \frac{2}{5}$$

and $$P(B) = \frac{5}{5+3} = \frac{5}{8} \quad \Rightarrow \quad P(\bar{B}) = 1 - \frac{5}{8} = \frac{3}{8}$$

The event E that X and Y contradict each other on an identical point can happen in the following mutually exclusive ways:

(i) X speaks the truth and Y tells a lie, *i.e.*, the event $A \cap \bar{B}$ happens,

(ii) X tells a lie and Y speaks the truth, *i.e.*, the event $\bar{A} \cap B$ happens.

Hence by addition theorem of probability the required probability is given by:

$$P(E) = P(i) + P(ii) = P(A \cap \bar{B}) + P(\bar{A} \cap B)$$

$$= P(A) \cdot P(\bar{B}) + P(\bar{A}) \cdot P(B),$$

[Since A and B are independent]

$$= \frac{3}{5} \times \frac{3}{8} + \frac{2}{5} \times \frac{5}{8} = \frac{19}{40} = 0{\cdot}475$$

Hence A and B are likely to contradict each other on an identical point in 47·5% of the cases.

Example 4·27. *A special dice is prepared such that the probabilities of throwing 1, 2, 3, 4, 5 and 6 points are :*

$$\frac{1-k}{6}, \frac{1+2k}{6}, \frac{1-k}{6}; \quad \frac{1+k}{6}, \frac{1-2k}{6}, \text{ and } \frac{1+k}{6}$$

respectively. If two such dice are thrown, find the probability of getting a sum equal to 9. **[Delhi Univ. B.Sc. (Stat. Hons.), 1988]**

Solution. Let (x, y) denote the numbers obtained in a thrown of two dice, x denoting the number on the first dice and y denoting the number on the second dice. The sum $S = x+y = 9$, can be obtained in the following mutually disjoint ways:

(i) (3, 6), *(ii)* (6, 3), *(iii)* (4, 5), *(iv)* (5, 4)

Hence by addition theorem of probability:

$$P(S=9) = P(3,6) + P(6,3) + P(4,5) + P(5,4)$$

$$= P(x=3)\,P(y=6) + P(x=6)\,P(y=3) + P(x=4)\,P(y=5) + P(x=5)\,P(y=4),$$

since the number on one dice is independent of the number on the other dice.

$$\therefore \quad P(S=9) = \frac{(1-k)}{6}\cdot\frac{(1+k)}{6} + \frac{(1+k)}{6}\cdot\frac{(1-k)}{6} + \frac{(1+k)}{6}\cdot\frac{(1-2k)}{6} + \frac{(1-2k)}{6}\cdot\frac{(1+k)}{6}$$

$$= 2\left(\frac{1+k}{36}\right)[(1-k)+(1-2k)]$$

$$= \frac{1}{18}(1+k)(2-3k)$$

Example 4·28. *(a) A and B alternately cut a pack of cards and the pack is shuffled after each cut. If A starts and the game is continued until one cuts a diamond, what are the respective chances of A and B first cutting a diamond?*

(b) One shot is fired from each of the three guns. E_1, E_2, E_3 denote the events that the target is hit by the first, second and third gun respectively. If $P(E_1) = 0·5$, $P(E_2) = 0·6$ and $P(E_3) = 0·8$ and E_1, E_2, E_3 are independent events, find the probability that (a) exactly one hit is registered, (b) at least two hits are registered.

Solution. *(a)* Let E_1 and E_2 denote the events of A and B cutting a diamond respectively. Then

$$P(E_1) = P(E_2) = \frac{13}{52} = \frac{1}{4} \Rightarrow P(\overline{E_1}) = P(\overline{E_2}) = \frac{3}{4}$$

If A starts the game, he can first cut the diamond in the following mutually exclusive ways:

(i) E_1 happens, *(ii)* $\overline{E_1} \cap \overline{E_2} \cap E_1$ happens, *(iii)* $\overline{E_1} \cap \overline{E_2} \cap \overline{E_1} \cap \overline{E_2} \cap E_1$ happens, and so on. Hence by addition theorem of probability, the probability 'p' that A first wins is given by

$$p = P(i) + P(ii) + P(iii) + \ldots\ldots$$

$$= P(E_1) + P(\overline{E_1} \cap \overline{E_2} \cap E_1) + P(\overline{E_1} \cap \overline{E_2} \cap \overline{E_1} \cap \overline{E_2} \cap E_1) + \ldots$$

$$= P(E_1) + P(\overline{E_1})\,P(\overline{E_2})\,P(E_1) + P(\overline{E_1})\cdot P(\overline{E_2})\,P(\overline{E_1})\,P(\overline{E_2})\,P(E_1) + \ldots$$

(By Compound Probability Theorem)

$$= \frac{1}{4} + \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4} + \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4} + \ldots\ldots$$

$$= \frac{\frac{1}{4}}{1-\frac{9}{16}} = \frac{4}{7}$$

The probability that B first cuts a diamond

$$= 1 - p = 1 - \frac{4}{7} = \frac{3}{7}$$

(b) We are given

$$P(\overline{E_1}) = 0·5,\ P(\overline{E_2}) = 0·4 \text{ and } P(\overline{E_3}) = 0·2$$

(a) Exactly one hit can be registered in the following mutually exclusive ways: *(i)* $E_1 \cap \overline{E_2} \cap \overline{E_3}$ happens, *(ii)* $\overline{E_1} \cap E_2 \cap \overline{E_3}$ happens, *(iii)* $\overline{E_1} \cap \overline{E_2} \cap E_3$ happens.

Hence by addition probability theorem, the required probability 'p' is given by :

$$p = P(E_1 \cap \overline{E_2} \cap \overline{E_3}) + P(\overline{E_1} \cap E_2 \cap \overline{E_3}) + P(\overline{E_1} \cap \overline{E_2} \cap E_3)$$

$$= P(E_1)\, P(\overline{E_2})\, P(\overline{E_3}) + P(\overline{E_1})\, P(E_2)\, P(\overline{E_3}) + P(\overline{E_1})\, P(\overline{E_2})\, P(E_3)$$

(Since E_1, E_2 and E_3 are independent)

$$= 0{\cdot}5 \times 0{\cdot}4 \times 0{\cdot}2 + 0{\cdot}5 \times 0{\cdot}6 \times 0{\cdot}2 + 0{\cdot}5 \times 0{\cdot}4 \times 0{\cdot}8 = 0{\cdot}26.$$

(b) At least two hits can be registered in the following mutually exclusive ways:

(i) $E_1 \cap E_2 \cap \overline{E_3}$ happens *(ii)* $E_1 \cap \overline{E_2} \cap E_3$ happens, *(iii)* $\overline{E_1} \cap E_2 \cap E_3$ happens. *(iv)* $E_1 \cap E_2 \cap E_3$ happens.

Required probability

$$= P(E_1 \cap E_2 \cap \overline{E_3}) + P(E_1 \cap \overline{E_2} \cap E_3) + P(\overline{E_1} \cap E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3)$$

$$= 0{\cdot}5 \times 0{\cdot}6 \times 0{\cdot}2 + 0{\cdot}5 \times 0{\cdot}4 \times 0{\cdot}8 + 0{\cdot}5 \times 0{\cdot}6 \times 0{\cdot}8 + 0{\cdot}5 \times 0{\cdot}6 \times 0{\cdot}8$$

$$= 0{\cdot}06 + 0{\cdot}16 + 0{\cdot}24 + 0{\cdot}24 = 0{\cdot}70$$

Example 4·29. *Three groups of children contain respectively 3 girls and 1 boy, 2 girls and 2 boys, and 1 girl and 3 boys. One child is selected at random from each group. Show that the chance that the three selected consist of 1 girl and 2 boys is 13/32.* **[Madurai Univ. B.Sc.,1988; Nagpur Univ. B.Sc.,1991]**

Solution. The required event of getting 1 girl and 2 boys among the three selected children can materialise in the following three mutually disjoint cases:

Group No. →	I	II	III
(i)	Girl	Boy	Boy
(ii)	Boy	Girl	Boy
(iii)	Boy	Boy	Girl

Hence by addition theorem of probability,

$$\text{Required probability} = P\,(i) + P\,(ii) + P\,(iii) \qquad \ldots(*)$$

Since the probability of selecting a girl from the first group is 3/4, of selecting a boy from the second is 2/4, and of selecting a boy from the third group is 3/4, and since these three events of selecting children from three groups are independent of each other, by compound probability theorem, we have

$$P\,(i) = \frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{9}{32}$$

Similarly, we have

$$P\,(ii) = \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{3}{32}$$

and
$$P\,(iii) = \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{1}{32}$$

Substituting in (*), we get

Required probability $= \frac{9}{32} + \frac{3}{32} + \frac{1}{32} = \frac{13}{32}$

EXERCISE 4 (b)

1. *(a)* Which function defines a probability space on $S = (e_1, e_2, e_3)$

(i) $P(e_1) = \frac{1}{4}, P(e_2) = \frac{1}{3}, P(e_3) = \frac{1}{2}$

(ii) $P(e_1) = \frac{2}{3}, P(e_2) = -\frac{1}{3}, P(e_3) = \frac{2}{3}$

(iii) $P(e_1) = \frac{1}{4}, P(e_2) = \frac{1}{3}, P(e_3) = \frac{1}{2}$, and

(iv) $P(e_1) = 0, P(e_2) = \frac{1}{3}, P(e_3) = \frac{2}{3}$

Ans. *(i)* No, *(ii)* No, *(iii)* No, and *(iv)* Yes

(b) Let $S = (e_1, e_2, e_3, e_4)$, and let P be a probability function on S.

(i) Find $P(e_1)$, if $P(e_2) = \frac{1}{3}, P(e_3) = \frac{1}{6}, P(e_4) = \frac{1}{9}$,

(ii) Find $P(e_1)$ and $P(e_2)$ if $P(e_3) = P(e_4) = \frac{1}{4}$ and $P(e_1) = 2P(e_2)$, and

(iii) Find $P(e_1)$ if $P[(e_2, e_3)] = \frac{2}{3}, P[(e_2, e_4)] = \frac{1}{2}$ and $P(e_2) = \frac{1}{3}$.

Ans. *(i)* $P(e_1) = \frac{7}{18}$, *(ii)* $P(e_1) = \frac{1}{3}$, $P(e_2) = \frac{1}{6}$, and *(iii)* $P(e_1) = \frac{1}{6}$

2. *(a)* With usual notations, prove that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Deduce a similar result for $P(A \cup B \cup C)$, where C is one more event.

(b) For any event : E_i, $P(E_i) = p_i$, $(i = 1, 2, 3)$; $P(E_i \cap E_j) = p_{ij}$, $(i, j = 1, 2, 3)$ and $P(E_1 \cap E_2 \cap E_3) = p_{123}$, find the probability that of the three events, *(i)* at least one, and *(ii)* exactly one happens.

(c) Discuss briefly the axiomatic approach to probability, illustrating by examples how it meets the deficiencies of the classical approach.

(d) If A and B are any two events, state the results giving

(i) $P(A \cup B)$ and *(ii)* $P(A \cap B)$.

A and B are mutually exclusive events and $P(A) = \frac{1}{2}, P(B) = \frac{1}{3}$. Find $P(A \cup B)$ and $P(A \cap B)$.

3. Let $S = \left\{1, \frac{1}{2}, \left(\frac{1}{2}\right)^2, \ldots, \left(\frac{1}{2}\right)^n\right\}$, be a classical event space and A, B be events given by

$$A = \left\{1, \frac{1}{2}\right\}, \quad B = \left\{\left(\frac{1}{2}\right)^k \mid k \text{ is an even positive integer}\right\}$$

Find $P(\bar{A} \cap \bar{B})$ [Calcutta Univ. B.Sc. (Stat Hons.), 1986]

4. What is a 'probability space'? State *(i)* the 'law of total probability' and *(ii)* Boole's inequality for events not necessarily mutually exclusive.

5. *(a)* Explain the following with examples:

(i) random experiment, *(ii)* an event, *(iii)* an event space. State the axioms of probability and explain their frequency interpretations.

A man forgets the last digit of a telephone number, and dials the last digit at random. What is the probability of calling no more than three wrong numbers?

(b) Define conditional probability and give its frequency interpretation. Show that conditional probabilities satisfy the axioms of probability.

6. Prove the following laws, in each case assuming the conditional probabilities being defined.

(a) $P(E \mid E) = 1$, *(b)* $P(\phi \mid F) = 0$

(c) If $E_1 \subseteq E_2$, then $E(E_1 \mid F) < P(E_2 \mid F)$

(d) $P(\bar{E} \mid F) = 1 - P(E \mid F)$

(e) $P(E_1 \cup E_2 \mid F) = P(E_1 \mid F) + P(E_2 \mid F) - P(P(E_1 \cap E_2 \mid F)$

(f) If $P(F) = 1$ then $P(E \mid F) = P(E)$

(g) $P(E - F) = P(E) - P(E \cap F)$

(h) If $P(F) > 0$, and E and F are mutully exclusive then $P(E \mid F) = 0$

(i) If $P(E \mid F) = P(E)$, then $P(E \mid \bar{F}) = P(E)$ and $P(\bar{E} \mid F) = P(\bar{E})$

7. *(a)* If $P(\bar{A}) = a$, $P(\bar{B}) = b$, then prove that $P(A \cap B) \geq 1 - a - b$.

(b) If $P(A) = \alpha$, $P(B) = \beta$, then prove that $P(A \mid B) \geq (\alpha + \beta - 1)/\beta$.

Hint. In each case use $P(A \cup B) \leq 1$

8. Prove or disprove:

(a) *(i)* If $P(A \mid B) \geq P(A)$, then $P(B \mid A) \geq P(B)$

(ii) If $P(A) = P(\bar{B})$, then $A = \bar{B}$.

[Delhi Univ. B.Sc. (Maths Hons.), 1988]

(b) If $P(A) = 0$, then $A = \phi$

[Delhi Univ. B.Sc. (Maths Hons.), 1990]

Ans. Wrong.

(c) For possibie events A, B, C,

(i) If $P(A) > P(B)$, then $P(A \mid C) > P(B \mid C)$

(ii) If $P(A \mid C) \geq P(B \mid C)$ and $P(A \mid \bar{C}) \geq P(B \mid \bar{C})$, then $P(A) \geq P(B)$. [Delhi Univ. B.Sc.(Maths Hons),1989]

(d) If $P(A) = 0$, then $P(A \cap B) = 0$.

[Delhi Univ. B.Sc. (Maths Hons.), 1986]

(e) *(i)* If $P(A) = P(B) = p$, then $P(A \cap B) \leq p^2$

(ii) If $P(B \mid \bar{A}) = P(B \mid A)$, then A and B are independent.

[Delhi Univ. B.Sc. (Maths Hons.), 1990]

(f) If $P(A) > 0, P(B) > 0$ and $P(A \mid B) = P(B \mid A)$, then $P(A) = P(B)$.

9. (a) Let A and B be two events, neither of which has probability zero. Then if A and B are disjoint, A and B are independent.

[Delhi Univ. B.Sc.(Stat. Hons.), 1986]

(b) Under what conditions does the following equality hold?

$$P(A) = P(A \mid B) + P(A \mid \overline{B})$$

[Punjab Univ. B.Sc. (Maths Hons.), 1992]

Ans. $B = S$ or $\overline{B} = S$

10. (a) If A and B are two events and the probability $P(B) \neq 1$, prove that

$$P(A \mid \overline{B}) = \frac{[P(A) - P(A \cap B)]}{[1 - P(B)]}$$

where $\overline{B}$ denotes the event complementary to B and hence deduce that

$$P(A \cap B) \geq P(A) + P(B) - 1$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

Also show that $P(A) >$ or $< P(A \mid B)$ according as

$$P(A \mid \overline{B}) > \text{ or } < P(A).$$

[Sri Venkat. Univ. B.Sc. 1992 ; Karnatak Univ. B.Sc.1991]

Hint. (i)

$$P(A \mid \overline{B}) = \frac{P(A \cap \overline{B})}{P(\overline{B})} = \frac{[P(A) - P(A \cap B)]}{[1 - P(B)]}$$

(ii) Since $P(A \mid \overline{B}) \leq 1$, $P(A) - P(A \cap B) \leq 1 - P(B)$

$\Rightarrow \quad P(A) + P(B) - 1 \leq P(A \cap B)$

(iii) $$\frac{P(A \mid \overline{B})}{P(A)} = \frac{P(\overline{B} \mid A)}{P(\overline{B})} = \frac{1 - P(B \mid A)}{1 - P(B)}$$

Now $\quad P(A \mid \overline{B}) > P(A)$ if $\{1 - P(B \mid A)\} > \{1 - P(B)\}$

i.e., if $\quad P(B \mid A) < P(B)$

i.e., if $\quad \dfrac{P(B \mid A)}{P(B)} < 1$

i.e., if $\quad \dfrac{P(A \mid B)}{P(A)} < 1$ i.e., if $P(A) > P(A \mid B)$

(b) If A and B are two mutually exclusive events show that

$$P(A \mid \overline{B}) = P(A)/[1 - P(B)]$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

(c) If A and B are two mutually exclusive events and $P(A \cup B) \neq 0$, then

$$P(A \mid A \cup B) = \frac{P(A)}{P(A) + P(B)}$$ [Guahati Univ. B.Sc. 1991]

(d) If A and B are two independent events show that

$$P(A \cup B) = 1 - P(\overline{A})P(\overline{B})$$

(e) If $\overline{A}$ denotes the non–occurrence of A, then prove that

$$P(\overline{A}_1 \cup \overline{A}_2 \cup \overline{A}_3) = 1 - P(A_1)\ P(A_2 \mid A_1)\ P(A_3 \mid A_1 \cap A_2)$$

[Agra Univ. B.Sc., 1987]

11. If A, B and C are three arbitrary events and

$$S_1 = P(A) + P(B) + P(C)$$
$$S_2 = P(A \cap B) + P(B \cap C) + P(C \cap A)$$
$$S_3 = P(A \cap B \cap C).$$

Prove that the probability that exactly one of the three events occurs is given by $S_1 - 2S_2 + 3S_3$.

12. (a) For the events $A_1, A_2, \ldots, A_n$ assuming

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i), \text{ prove that}$$

(i) $P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(\overline{A}_i)$ and that

(ii) $P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$

[Sardar Patel Univ. B.Sc. Nov.1992]

(b) Let A, B and C denote events. If $P(A \mid C) \geq P(B \mid C)$ and $P(A \mid \overline{C}) \geq P(B \mid \overline{C})$, then show that $P(A) \geq P(B)$.

[Calcutta Univ. B.Sc. (Maths Hons.), 1992]

13. (a) If A and B are independent events defined on a given probability space $(\Omega, \mathcal{A}, P(.))$, then prove that A and $\overline{B}$ are independent, $\overline{A}$ and $\overline{B}$ are independent. [Delhi Univ. B.Sc. (Maths Hons.), 1988]

(b) A, B and C are three events such that A and B are independent, $P(C) = 0$. Show that A, B and C are independent.

(c) An event A is known to be independent of the events $B, B \cup C$ and $B \cap C$. Show that it is also independent of C. [Nagpur Univ. B.Sc.1992]

(d) Show that if an event C is independent of two mutually exclusive events A and B, then C is also independent of $A \cup B$.

(e) The outcome of an experiment is equally likely to be one of the four points in three–dimensional space with rectangular coordinates $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ and $(1, 1, 1)$. Let E, F and G be the events : x–coordinate=1, y–coordinate=1 and z–coordinate=1, respectively. Check if the events E, F and G are independent. (Calcutta Univ. B.Sc., 1988)

14. Explain what is meant by "Probability Space". You fire at a target with each of the three guns; A, B and C denote respectively the event — hit the target with the first, second and third gun. Assuming that the events are independent and have probabilities $P(A) = a$, $P(B) = b$ and $P(C) = c$, express in terms of A, B and C the following events:

(i) You will not hit the target at all.

(*ii*) You will hit the target at least twice. Find also the probabilities of these events. **[Sardar Patel Univ. B.Sc., 1990)**

15. (*a*) Suppose A and B are any two events and that $P(A) = p_1$, $P(B) = p_2$ and $P(A \cap B) = p_3$. Show that the formula of each of the following probabilities in terms of p_1, p_2 and p_3 can be expressed as follows :

(*i*) $P(\overline{A} \cup \overline{B}) = 1 - p_2$ (*ii*) $P(\overline{A} \cap \overline{B}) = 1 - p_1 - p_2 + p_3$

(*iii*) $P(A \cap \overline{B}) = p_1 - p_3$ (*iv*) $P(\overline{A} \cap B) = p_2 - p_3$

(*v*) $P(\overline{A \cap B}) = 1 - p_3$ (*vi*) $P(\overline{A} \cup B) = 1 - p_1 + p_3$

(*vii*) $P(\overline{A \cup B}) = 1 - p_1 - p_2 + p_3$ (*viii*) $P[\overline{A} \cap (A \cup B)] = p_2 - p_3$

(*ix*) $P[A \cup (\overline{A} \cap B)] = p_1 + p_2 - p_3$

(*x*) $P(A \mid B) = \dfrac{p_3}{p_2}$ and $P(B \mid A) = \dfrac{p_3}{p_1}$

(*xi*) $P(\overline{A} \mid \overline{B}) = \dfrac{1 - p_1 - p_2 + p_3}{1 - p_2}$ and $P(\overline{B} \mid \overline{A}) = \dfrac{1 - p_1 - p_2 + p_3}{1 - p_1}$

[Allahabad Univ. B.Sc. (Stat.), 1991]

(*b*) If $P(A) = 1/3$, $P(B) = 3/4$ and $P(A \cup B) = 11/12$, find $P(A \mid B)$ and $P(B \mid A)$.

(*c*) Let $P(A) = p$, $P(A \mid B) = q$, $P(B \mid A) = r$. Find the relation between the numbers p, q and r such that $\overline{A}$ and $\overline{B}$ are mutually exclusive.

[Delhi Univ. B.Sc. (Maths Hons.), 1985]

Hint. $P(AB) = P(A)\, P(B \mid A) = P(B) \,.\, P(A \mid B)$

$\Rightarrow$ $P(AB) = pr = P(B).\, q$ $\Rightarrow$ $P(B) = pr/q$

If $\overline{A}$ and $\overline{B}$ are mutually disjoint, then $P(\overline{A} \cap \overline{B}) = 0$.

$\Rightarrow$ $1 - P(A \cup B) = 0 \Rightarrow 1 - [p + (pr/q) - pr] = 0$

16. (*a*) In terms of probabilities, $p_1 = P(A)$, $p_2 = P(B)$ and $p_3 = P(A \cap B)$;

Express (*i*) $P(A \cup B)$, (*ii*) $P(A \mid B)$, (*iii*) $P(\overline{A} \cap B)$ under the condition that (*i*) A and B are mutually exclusive, (*ii*) A and B are mutually independent.

(*b*) Let A and B be the possible outcomes of an experiment and suppose

$P(A) = 0{\cdot}4$, $P(A \cup B) = 0{\cdot}7$ and $P(B) = p$

(*i*) For what choice of p are A and B mutually exclusive ?

(*ii*) For what choice of p are A and B independent ?

[Aligarh Univ. B.Sc., 1988 ; Guwahati Univ. B.Sc., 1991]

Ans. (*i*) 0·3, (*ii*) 0·5

(*c*) Let A_1, A_2, A_3, A_4 be four independent events for which $P(A_1) = p$, $p(A_2) = q$, $P(A_3) = r$ and $P(A_4) = s$. Find the probability that

(*i*) at least one of the events occurs, (*ii*) exactly two of the events occur, and (*iii*) at most three of the events occur. **[Civil Services (Main), 1985]**

17. (*a*) Two six-faced unbiased dice are thrown. Find the probability that the sum of the numbers shown is 7 or their product is 12.

Ans. 2/9

(b) Defects are classifed as A, B or C, and the following probabilities have been determined from available production data :

$P(A) = 0{\cdot}20$, $P(B) = 0{\cdot}16$, $P(C) = 0{\cdot}14$, $P(A \cap B) = 0{\cdot}08$, $P(A \cap C) = 0{\cdot}05$, $P(B \cap C) = 0{\cdot}04$, and $P(A \cap B \cap C) = 0{\cdot}02$.

What is the probability that a randomly selected item of product will exhibit at least one type of defect ? What is the probability that it exhibits both A and B defects but is free from type C defect ? **[Bombay Univ. B.Sc., 1991]**

(c) A language class has only three students A, B, C and they independently attend the class. The probabilities of attendance of A, B and C on any given day are $1/2$, $2/3$ and $3/4$ respectively. Find the probability that the total number of attendances in two consecutive days is exactly three.

[Lucknow Univ. B.Sc. 1990; Calcutta Univ. B.Sc.(Maths Hons.), 1986]

18. *(a)* Cards are drawn one by one from a full deck. What is the probability that exactly 10 cards will precede the first ace. **[Delhi Univ. B.Sc.,1988]**

Ans. $\left(\frac{48}{52} \times \frac{47}{51} \times \frac{46}{50} \times \ldots \times \frac{39}{43}\right) \times \frac{4}{42} = \frac{164}{4165}$

(b) Each of two persons tosses three fair coins. What is the probability that they obtain the same number of heads.

Ans. $\left(\frac{1}{8}\right)^2 + \left(\frac{3}{8}\right)^2 + \left(\frac{3}{8}\right)^2 + \left(\frac{1}{8}\right)^2 = \frac{5}{16}$.

19. *(a)* Given that A, B and C are mutually exclusive events, explain why each of the following is not a permissible assignment of probabilities.

(i) $P(A) = 0{\cdot}24$, $P(B) = 0{\cdot}4$ and $P(A \cup C) = 0{\cdot}2$,

(ii) $P(A) = 0{\cdot}7$, $P(B) = 0{\cdot}1$ and $P(B \cap C) = 0{\cdot}3$

(iii) $P(A) = 0{\cdot}6$, $P(A \cap \overline{B}) = 0{\cdot}5$

(b) Prove that for n arbitrary independent events $A_1, A_2, \ldots, A_n$

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) + P(\overline{A_1})\, P(\overline{A_2}) \ldots P(\overline{A_n}) = 1.$$

(c) $A_1, A_2, \ldots, A_n$ are n independent events with

$$P(A_i) = 1 - \frac{1}{\alpha^i}, \; i = 1, 2, \ldots, n.$$

Find the value of $P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n)$. **(Nagpur Univ. B.Sc., 1987)**

Ans. $1 - \frac{1}{\alpha^{n(n+1)/2}}$

(d) Suppose the events $A_1, A_2, \ldots, A_n$ are independent and that

$P(A_i) = \frac{1}{i+1}$ for $1 \le i \le n$. Find the probability that none of the n events occurs, justifying each step in your calculations.

Ans. $1/(n+1)$

20. *(a)* A denotes getting a heart card, B denotes getting a face card (King, Queem or Jack), $\overline{A}$ and $\overline{B}$ denote the complementary events. A card is drawn at random from a full deck. Compute the following probabilities.

(*i*) $P(A)$, (*ii*) $P(A \cap \overline{B})$, (*iii*) $P(A \cup B)$, (*iv*) $P(A \cap B)$,

(*v*) $P(\overline{A} \cup B)$.

Assume natural assignment of probabilities.

Ans. (*i*) 1/4, (*ii*) 5/26, (*iii*) 11/26, (*iv*) 3/5, (*v*) 21/26.

(*b*) A town has two doctors X and Y operating independently. If the probability that doctor X is available is 0·9 and that for Y is 0·8, what is the probability that at least one doctor is available when needed? **[Gorakhpur Univ. B.Sc., 1988]**

Ans. 0·98

21. (*a*) The odds that a book will be favourably reviewed by 3 independent critics are 5 to 2, 4 to 3 and 3 to 4 respectively. What is the probability that, of the three reviews, a majority will be favourable? **[Gauhati Univ. BSc., 1987]**

Ans. 209/343.

(*b*) A, B and C are independent witnesses of an event which is known to have occurred. A speaks the truth three times out of four, B four times out of five and C five times out of six. What is the probability that the occurrence will be reported truthfully by majority of three witnesses?

Ans. 31/60.

(*c*) A man seeks advice regarding one of two possible courses of action from three advisers who arrived at their recommendations independently. He follows the recommendation of the majority. The probability that the individual advisers are wrong are 0·1, 0·05 and 0·05 respectively. What is the probability that the man takes incorrect advise ? **[Gujarat Univ. B.Sc., 1987]**

22. (*a*) The odds against a certain event are 5 to 2 and odds in favour of another (independent) event are 6 to 5. Find the chance that at least one of the events will happen. **(Madras Univ. B.Sc.,1987)**

Ans. 52/77.

(*b*) A person takes four tests in succession. The probability of his passing the first test is p, that of his passing each succeeding test is p or $p/2$ according as he passes or fails the preceding one. He qualifies provided he passes at least three tests. What is his chance of qualifying. **[Gauhati Univ. B.Sc. (Hons.) 1988]**

23. (*a*) The probability that a 50-years old man will be alive at 60 is 0·83 and the probability that a 45-years old woman will be alive at 55 is 0·87. What is the probability that a man who is 50 and his wife who is 45 will both be alive 10 years hence?

Ans. 0·7221.

(*b*) It is 8:5 against a husband who is 55 years old living till he is 75 and 4:3 against his wife who is now 48, living till she is 68. Find the probability that (*i*) the couple will be alive 20 years hence, and (*ii*) at least one of them will be alive 20 years hence.

Ans (*i*) 15/91, (*ii*) 59/91.

(c) A husband and wife appear in an interview for two vacancies in the same post. The probability of husband's selection is 1/7 and that of wife's selection is 1/5. What is the probability that only one of them will be selected ?

Ans. 2/7 **[Delhi Univ. B.Sc., 1986]**

24. *(a)* The chances of winning of two race-horses are 1/3 and 1/6 respectively. What is the probability that at least one will win when the horses are running *(a)* in different races, and *(b)* in the same race?

Ans. *(a)* 8/18 *(b)* 1/2

(b) A problem in statistics is given to three students whose chances of solving it are 1/2, 1/3 and 1/4. What is the probability that the problem will be solved?

Ans. 3/4 **[Meerut Univ. B.Sc., 1990]**

25. *(a)* Ten pairs of shoes are in a closet. Four shoes are selected at random. Find the probability that there will be at least one pair among the four shoes selected?

Ans. $1-\dfrac{{}^{10}C_4 \times 2^4}{{}^{20}C_4}$

(b) From 100 tickets numbered 1, 2, ..., 100 four are drawn at random. What is the probability that 3 of them will bear number from 1 to 20 and the fourth will bear any number from 21 to 100 ?

Ans. $\dfrac{{}^{20}C_3 \times {}^{80}C_1}{{}^{100}C_4}$

26. A six faced die is so biased that it is twice as likely to show an even number as an odd number when thrown. It is thrown twice. What is the probability that the sum of the two numbers thrown is odd?

Ans. 4/9

27. From a group of 8 children, 5 boys and 3 girls, three children are selected at random. Calculate the probabilities that selected group contains *(i)* no girl, *(ii)* only one girl, *(iii)* one particular girl, *(iv)* at least one girl, and *(v)* more girls than boys.

Ans. *(i)* 5/28, *(ii)* 15/28, *(iii)* 5/28, *(iv)* 23/28, *(v)* 2/7.

28. If three persons, selected at random, are stopped on a street, what are the probabilities that :

(a) all were born on a Friday;

(b) two were born on a Friday and the other on a Tuesday;

(c) none was born on a Monday.

Ans. *(a)* 1/343, *(b)* 3/343, *(c)* 216/343.

29. *(a)* A and B toss a coin alternately on the understanding that the first who obtains the head wins. If A starts, show that their respective chances of winning are 2/3 and 1/3.

(b) A, B and C, in order, toss a coin. The first one who throws a head wins. If A starts, find their respectivre chances of winning. (Assume that the game may

continue indefinitely.)

Ans. 4/7, 2/7, 1/7.

(c) A man alternately tosses a coin and throws a die, beginning with coin. What is the probability that he will get a head before he gets a '5 or 6' on die?

Ans. 3/4.

30. *(a)* Two ordinary six-sided dice are tossed.

(i) What is the probability that both the dice show the number 5.

(ii) What is the probability that both the dice show the same number.

(iii) Given that the sum of two numbers shown is 8, find the conditional probability that the number noted on the first dice is larger than the number noted on the second dice.

(b) Six dice are thrown simultaneously. What is the probability that all will show different faces?

31. *(a)* A bag contains 10 balls, two of which are red, three blue and five black. Three balls are drawn at random from the bag, that is every ball has an equal chance of being included in the three. What is the probability that

(i) the three balls are of different colours,

(ii) two balls are of the same colour, and

(iii) the balls are all of the same colour?

Ans. *(i)* 30/120, *(ii)* 79/120, *(iii)* 11/120.

(b) A is one of six horses entered for a race and is to be ridden by one of the two jockeys B and C. It is 2 to 1 that B rides A, in which case all the horses are equally likely to win, with rider C, A's chance is trebled.

(i) Find the probability that A wins.

(ii) What are odds against A's winning?

[Shivaji Univ. B.Sc. (Stat. Hons.), 1992]

Hint. Probability of A's winning

$$= P\,(B \text{ rides } A \text{ and } A \text{ wins}) + P\,(C \text{ rides } A \text{ and } A \text{ wins})$$

$$= \frac{2}{3} \times \frac{1}{6} + \frac{1}{3} \times \frac{3}{6} = \frac{5}{18}$$

$\therefore$ Probability of A's losing $= 1 - 5/18 = 13/18$.

Hence odds against A's winning are : $13/18 : 5/18$, *i.e.*, $13 : 5$.

32. *(a)* Two-third of the students in a class are boys and the rest girls. It is known that the probability of a girl getting a first class is 0·25 and that of boy getting a first class is 0·28. Find the probability that a student chosen at random will get first class marks in the subject.

Ans. 0·27

(b) You need four eggs to make omelettes for breakfast. You find a dozen eggs in the refrigerator but do not realise that two of these are rotten. What is the probability that of the four eggs you choose at random

(i) none is rotten,

(*ii*) exactly one is rotten?

Ans. (*i*) 625/1296 ; (*ii*) 500/1296.

(*c*) The probability of occurrence of an event A is 0·7, the probability of non-occurrence of another event B is 0·5 and that of at least one of A or B not occurring is 0·6. Find the probability that at least one of A or B occurs.

[Mysore Univ. B.Sc., 1991]

33. (*a*) The odds against A solving a certain problem are 4 to 3 and odds in favour of B solving the same problem are 7 to 5. What is the probability that the problem is solved if they both try independently? **[Gujarat Univ.B.Sc., 1987]**

Ans. 16/21

(*b*) A certain drug manufactured by a company is tested chemically for its toxic nature. Let the event 'the drug is toxic' be denoted by E and the event 'the chemical test reveals that the drug is toxic' be denoted by F. Let $P(E) = \theta$, $P(F \mid E) = P(\bar{F} \mid \bar{E}) = 1 - \theta$. Then show that probability that the drug is not toxic given that the chemical test reveals that it is toxic is free from θ.

Ans. 1/2 **[M.S. Baroda Univ. B.Sc., 1992]**

34. A bag contains 6 white and 9 black balls. Four balls are drawn at a time. Find the probability for the first draw to give 4 white and the second draw to give 4 black balls in each of the following cases :

(*i*) The balls are replaced before the second draw.

(*ii*) The balls are not replaced before the second draw.

[Jammu Univ. B.Sc., 1992]

Ans. (*i*) $\dfrac{{}^6C_4}{{}^{15}C_4} \times \dfrac{{}^9C_4}{{}^{15}C_4}$ (*ii*) $\dfrac{{}^6C_4}{{}^{15}C_4} \times \dfrac{{}^9C_4}{{}^{11}C_4}$

35. The chances that doctor A will diagnose a disease X correctly is 60%. The chances that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by wrong diagnosis is 70%. A patient of doctor A, who had disease X, died. What is the chance that his disease was diagnosed correctly?

Hint. Let us define the following events:

E_1 : Disease X is diagnosed correctly by doctor A.

E_2 : A patient (of doctor A) who has disease X dies.

Then we are given :

$$P(E_1) = 0{\cdot}6 \quad \Rightarrow \quad P(\bar{E}_1) = 1 - 0{\cdot}6 = 0{\cdot}4$$

$$\text{and} \quad P(E_2 \mid E_1) = 0{\cdot}4 \quad \text{and} \quad P(E_2 \mid \bar{E}_1) = 0{\cdot}7$$

$$\text{We want } P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1 \cap E_2)}{P(E_1 \cap E_2) + P(\bar{E}_1 \cap E_2)} = \frac{6}{13}$$

36. The probability that at least 2 of 3 people A, B and C will survive for 10 years is 247/315. The probability that A alone will survive for 10 years is 4/105 and the probability that C alone will die within 10 years is 2/21. Assuming that the events of the survival of A, B and C can be regarded as independent, calculate the probability of surviving 10 years for each person.

Ans. 3/5, 5/7, 7/9.

37. A and B throw alternately a pair of unbiased dice, A beginning. A wins if he throws 7 before B throws 6, and B wins if he throws 6 before A throws 7. If A and B respectively denote the events that A wins and B wins the series, and a and b respectively denote the events that it is A's and B's turn to throw the dice, show that

$$(i)\ P(A \mid a) = \frac{1}{6} + \frac{5}{6} P(A \mid b), \quad (ii)\ P(A \mid b) = \frac{31}{36} P(A \mid a),$$

$$(iii)\ P(B \mid a) = \frac{5}{6} P(B \mid b), \text{ and } (iv)\ P(B \mid b) = \frac{5}{36} + \frac{13}{36} P(B \mid a),$$

Hence or otherwise, find $P(A \mid a)$ and $P(B \mid a)$. Also comment on the result that $P(A \mid a) + P(B \mid a) = 1$.

38. A bag contains an assortment of blue and red balls. If two balls are drawn at randon, the probability of drawing two red balls is five times the probability of drawing two blue balls. Furthermore, the probability of drawing one ball of each colour is six times the probabililty of drawing two blue balls. How many red and blue balls are there in the bag?

Hint. Let number of red and blue balls in the bag be r and b respectively. Then

$$p_1 = \text{Prob. of drawing two red balls} = \frac{r(r-1)}{(r+b)(r+b-1)}$$

$$p_2 = \text{Prob. of drawing two blue balls} = \frac{b(b-1)}{(r+b)(r+b-1)}$$

$$p_3 = \text{Prob. of drawing one red and one blue ball} = \left[\frac{2br}{(r+b)(r+b-1)}\right]$$

Now $\quad p_1 = 5p_2$ and $p_3 = 6p_2$

$\therefore \quad r(r-1) = 5b(b-1)$ and $2br = 6b(b-1)$

Hence $b = 3$ and $r = 6$.

39. Three newspapers A, B and C are published in a certain city. It is estimated from a survey that 20% read A; 16% read B, 14% read C, 8% read A and B, 5% read A and C, 4% read B and C and 2% read all the three newspapers. What is the probability that a normally chosen person

(i) does not read any paper, *(ii)* does not read C

(iii) reads A but not B, *(iv)* reads only one of these papers, and

(v) reads only two of these papers.

Ans. *(i)* 0·65, *(ii)* 0·86, *(iii)* 0·12, *(iv)* 0·22, *(v)* 0·11.

40. *(a)* A die is thrown twice, the event space S consisting of the 36 possible pairs of outcomes (a,b) each assigned probability 1/36. Let A, B and C denote the following events :

$A = \{(a,b) \mid a \text{ is odd}\}$, $B = \{(a,b) \mid b \text{ is odd}\}$, $C = \{(a,b) \mid a + b \text{ is odd.}\}$

Check whether A, B and C are independent or independent in pairs only.

[Calcutta Univ. B.Sc. Hons., 1985]

(b) Eight tickets numbered 111, 121, 122, 122, 211, 212, 212, 221 are placed in a hat and stirred. One of them is then drawn at random. Show that the event A : "the first digit on the ticket drawn will be 1", B : "the second digit on the ticket drawn will be 1," and C : "the third digit on the ticket drawn will be 1", are not pairwise independent although

$$P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$$

41. *(a)* Four identical marbles marked 1, 2, 3 and 123 respectively are put in an urn and one is drawn at random. Let A_i, $(i = 1, 2, 3)$, denote the event that the number i appears on the drawn marble. Prove that the events A_1, A_2 and A_3 are pairwise independent but not mutually independent.

[Gauhati Univ. B.Sc. (Hons.), 1988]

Hint. $P(A_1) = \frac{1}{2} = P(A_2) = P(A_3)$; $P(A_1 A_2) = P(A_1 A_3) = P(A_2 A_3) = \frac{1}{4}$

$P(A_1 A_2 A_3) = \frac{1}{4}$.

(b) Two fair dice are thrown independently. Define the following events :

A : Even number on the first dice

B : Even number on the second dice.

C : Same number on both dice.

Discuss the independence of the events A, B and C.

(c) A die is of the shape of a regular tetrahedron whose faces bear the numbers 111, 112, 121, 122. A_1, A_2, A_3 are respectively the events that the first two, the last two and the extreme two digits are the same, when the die is tossed at random. Find whether or not the events A_1, A_2, A_3 are *(i)* pairwise independent, *(ii)* mutually (*i.e.* completely) independent. Determine $P(A_1 \mid A_2 A_3)$) and explain its value by by argument. **[Civil Services (Main), 1983]**

42. *(a)* For two events A and B we have the following probabilities:

$$P(A) = P(A \mid B) = \frac{1}{4} \text{ and } P(B \mid A) = \frac{1}{2}.$$

Check whether the following statements are true or false :

(i) A and B are mutually exclusive, *(ii)* A and B are independent, *(iii)* A is a subevent of B, and *(iv)* $P(\bar{A} \mid B) = \frac{3}{4}$

Ans. *(i)* False, *(ii)* True,, *(iii)* False, and *(iv)* True.

(b) Consider two events A and B such that $P(A) = 1/4, P(B \mid A) = 1/2$, $P(A \mid B) = 1/4$. For each of the following statements, ascertain whether it is true or false :

(i) A is a sub-event of B, *(ii)* $P(\bar{A} \mid \bar{B}) = 3/4$,

(iii) $P(A \mid B) + P(A \mid \bar{B}) = 1$

43. *(a)* Let A and B be two events such that $P(A) = \frac{3}{4}$ and $P(B) = \frac{5}{8}$.

Show that

(i) $P(A \cup B) \geq \frac{3}{4}$, (ii) $\frac{3}{8} \leq P(A \cap B) \leq \frac{5}{8}$, and (iii) $\frac{1}{8} \leq P(A \cap \bar{B}) \leq \frac{3}{8}$.

[Coimbatore Univ. B.E., Nov. 1990; Delhi Univ. B.Sc.(Stat. Hons.),1986]

(b) Given two events A and B. If the odds against A are 2 to 1 and those in favour of $A \cup B$ are 3 to 1, show that

$$\frac{5}{12} \leq P(B) \leq \frac{3}{4}$$

Give an example in which $P(B) = 3/4$ and one in which $P(B) = 5/12$.

44. Let A and B be events, neither of which has probability zero. Prove or disprove the following events :

(i) If A and B are disjoint, A and B are independent,

(ii) If A and B are independent, A and B are disjoint.

45. *(a)* It is given that $P(A_1 \cup A_2) = \frac{5}{6}$, $P(A_1 \cap A_2) = \frac{1}{3}$ and $P(\bar{A}_2) = \frac{1}{2}$, where $P(\bar{A}_2)$ stands for the probability that A_2 does not happen. Determine $P(A_1)$ and $P(A_2)$.

Hence show that A_1 and A_2 are independent.

Ans. $P(A_1) = \frac{2}{3}$, $P(A_2) = \frac{1}{2}$

(b) A and B are events such that

$$P(A \cup B) = \frac{3}{4},\ P(A \cap B) = \frac{1}{4},\ \text{and}\ P(\bar{A}) = \frac{2}{3}.$$

Find *(i)* $P(A)$, *(ii)* $P(B)$ and *(iii)* $P(A \cap \bar{B})$.

(Madras Univ. B.E., 1989)

Ans. *(i)* 1/3, *(ii)* 2/3 *(iii)* 1/12.

46. A thief has a bunch of n keys, exactly one of which fits a lock. If the thief tries to open the lock by trying the keys at random, what is the probability that he requires exactly k attempts, if he rejects the keys already tried? Find the same probability if he does not reject the keys already tried.

(Aligarh Univ. B.Sc., 1991)

Ans. *(i)* $\frac{1}{n}$, *(ii)* $\left(\frac{n-1}{n}\right)^{k-1} \cdot \frac{1}{n}$

(b) There are M urns numbered 1 to M and M balls numbered 1 to M. The balls are inserted randomly in the urns with one ball in each urn. If a ball is put into the urn bearing the same number as the ball, a match is said to have occurred. Find the probability that no match has occurred. **[Civil Services (Main), 1984]**

Hint. See Example 4·54 page 4·97.

47. If n letters are placed at random in n correctly addressed envelopes, find the probability that

(i) none of the letters is placed in the correct envelope,

(ii) At least one letter goes to the correct envelope,
(iii) All letters go to the correct envelopes.

[Delhi Univ. B.Sc. (Stat Hons.), 1987, 1984]

48. An urn contains n white and m black balls, a second urn contains N white and M black balls. A ball is randomly transferred from the first to the second urn and then from the second to the first urn. If a ball is now selected randomly from the first urn, prove that the probability that it is white is

$$\frac{n}{n+m}+\frac{mN-nM}{(n+m)^2\,(N+M+1)}$$

[Delhi Univ. B.Sc. (Stat.Hons.) 1986]

Hint. Let us define the following events :

B_i : Drawing of a black ball from the ith urn, $i=1,2$.
W_i : Drawing of a white ball from the ith urn , $i=1,2$.

The four distinct possibilities for the first two exchanges are $B_1W_2, B_1B_2, W_1B_2, W_1W_2$. Hence if E denotes the event of drawing a white ball from the first urn after the exchanges, then

$$P(E)=P(B_1W_2E)+P(B_1B_2E)+P(W_1B_2E)+P(W_1W_2E) \qquad \ldots(*)$$

We have :

$$P(B_1W_2E)=P(B_1)\,.\,P(W_2\mid B_1)\,P(E\mid B_1W_2)=\frac{m}{m+n}\times\frac{N}{M+N+1}\times\frac{n+1}{m+n}$$

$$P(B_1B_2E)=P(B_1)\,.\,P(B_2\mid B_1)\,.\,P(E\mid B_1B_2)=\frac{m}{m+n}\times\frac{M+1}{M+N+1}\times\frac{n}{m+n}$$

$$P(W_1B_2E)=P(W_1)\,.\,P(B_2\mid W_1)\,.\,P(E\mid W_1B_2)=\frac{n}{m+n}\times\frac{M}{M+N+1}\times\frac{n-1}{m+n}$$

$$P(W_1W_2E)=P(W_1)\,.\,P(W_2\mid W_1)\,.\,P(E\mid W_1W_2)=\frac{n}{m+n}\times\frac{N+1}{M+N+1}\times\frac{n}{m+n}$$

Substituting in $(*)$ and simplifying we get the result.

49. A particular machine is prone to three similar types of faults A_1, A_2 and A_3. Past records on breakdowns of the machine show the following : the probability of a breakdown (*i.e.*, at least one fault) is 0·1; for each i, the probability that fault A_i occurs and the others do not is 0·02 ; for each pair i, j the probability that A_i and A_j occur but the third fault does not is 0·012. Determine the probabilities of

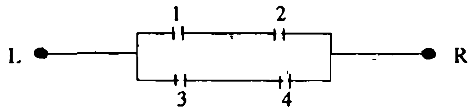
(a) the fault of type A_1 occurring irrespective of whether the other faults occur or not,

(b) a fault of type A_1 given that A_2 has occurred,

(c) faults of type A_1 and A_2 given that A_3 has occurred.

[London U. B.Sc. 1976]

50. The probability of the closing of each relay of the circuit shown below is given by p. If all the relays function independently, what is the probability that a circuit exists between the terminals L and R?



Ans. $p^2(2-p^2)$.

4·9. Bayes Theorem. *If $E_1, E_2, \ldots, E_n$ are mutually disjoint events with $P(E_i) \neq 0$, $(i = 1, 2, \ldots, n)$ then for any orbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have*

$$P(E_i \mid A) = \frac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)}, \quad i = 1, 2, \ldots, n. \qquad \ldots(4\cdot12)$$

Proof. Since $A \subset \bigcup_{i=1}^{n} E_i$, we have

$$A = A \cap \left(\bigcup_{i=1}^{n} E_i\right) = \bigcup_{i=1}^{n} (A \cap E_i) \qquad \text{[By distributive law]}$$

Since $(A \cap E_i) \subset E_i$, $(i = 1, 2, \ldots, n)$ are mutually disjoint events, we have by addition theorem of probability (or Axiom 3 of probability)

$$P(A) = P\left[\bigcup_{i=1}^{n} (A \cap E_i)\right] = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)\,P(A \mid E_i), \qquad \ldots(*)$$

by compound theorem of probability.

Also we have

$$P(A \cap E_i) = P(A)\,P(E_i \mid A)$$

$$P(E_i \mid A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)} \qquad \text{[From (*)]}$$

Remarks. 1. The probabilities $P(E_1), P(E_2), \ldots, P(E_n)$ are termed as the '*a priori probabilities*' because they exist before we gain any information from the experiment itself.

2. The probabilities $P(A \mid E_i)$, $i = 1, 2, \ldots, n$ are called '*likelihoods*' because they indicate how likely the event A under consideration is to occur, given each and every *a priori* probability.

3. The probabilities $P(E_i \mid A)$, $i = 1, 2, \ldots, n$ are called '*posterior probabilities*' because they are determined after the results of the experiment are known.

4. From (*) we get the following important result:

"If the events $E_1, E_2, \ldots, E_n$ constitute a partition of the sample space S and $P(E_i) \neq 0$, $i = 1, 2, \ldots, n$, then for any event A in S we have

$$P(A) = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)\, P(A \mid E_i) \qquad \ldots(4 \cdot 12\, a)$$

Cor. *(Bayes theorem for future events)*

The probability of the materialisation of another event C, given

$$P(C \mid A \cap E_1),\, P(C \mid A \cap E_2),\, \ldots,\, P(C \mid A \cap E_n) \quad \textit{is}$$

$$P(C \mid A) = \frac{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)\, P(C \mid E_i \cap A)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)} \qquad \ldots(4 \cdot 12\, b)$$

Proof. Since the occurrence of event A implies the occurrence of one and only one of the events $E_1, E_2, \ldots, E_n$, the event C (granted that A has occurred) can occu in the following mutually exclusive ways:

$$C \cap E_1, C \cap E_2, \ldots, C \cap E_n$$

i.e., $\quad C = (C \cap E_1) \cup (C \cap E_2) \cup \ldots \cup (C \cap E_n)$

$\Rightarrow \quad C \mid A = [(C \cap E_1) \mid A] \cup [(C \cap E_2) \mid A] \cup \ldots \cup [(C \cap E_n) \mid A]$

$\therefore \quad P(C \mid A) = P[(C \cap E_1) \mid A] + P[(C \cap E_2) \mid A] + \ldots + P[(C \cap E_n) \mid A]$

$$= \sum_{i=1}^{n} P[(C \cap E_i) \mid A]$$

$$= \sum_{i=1}^{n} P(E_i \mid A)\; P[C \mid (E_i \cap A)]$$

Substituting the value of $P(E_i \mid A)$ from (*), we get

$$P(C \mid A) = \frac{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)\; P(C \mid E_i \cap A)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)}$$

Remark. It may happen that the materialisation of the event E_i makes C independent of A, then we have

$$P(C \mid E_i \cap A) = P(C \mid E_i),$$

and the above formula reduces to

$$P(C \mid A) = \frac{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)\; P(C \mid E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)} \qquad \ldots(4 \cdot 12\, c)$$

The event C can be considered in regard to A, as Future Event.

Example 4·30. *In 1989 there were three candidates for the position of principal – Mr. Chatterji, Mr. Ayangar and Dr. Singh – whose chances of getting the appointment are in the proportion 4:2:3 respectively. The probability that Mr. Chatterji if selected would introduce co-education in the college is 0·3. The probabilities of Mr. Ayangar and Dr. Singh doing the same are respectively 0·5 and 0·8. What is the probability that there was co-education in the college in 1990?*

[Delhi Univ. B.Sc.(Stat. Hons.), 1992; Gorakhpur Univ. B.Sc., 1992]

Solution. Let the events and probabilities be defined as follows:

A : Introduction of co-education

E_1 : Mr. Chatterji is selected as principal

E_2 : Mr. Ayangar is selected as principal

E_3 : Dr. Singh is selected as principal.

Then

$$P(E_1) = \frac{4}{9}, \quad P(E_2) = \frac{2}{9} \quad \text{and} \quad P(E_3) = \frac{3}{9}$$

$$P(A \mid E_1) = \frac{3}{10}, \quad P(A \mid E_2) = \frac{5}{10} \quad \text{and} \quad P(A \mid E_3) = \frac{8}{10}$$

$$\begin{aligned} \therefore \quad P(A) &= P[(A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3)] \\ &= P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3) \\ &= P(E_1)P(A \mid E_1) + P(E_2)P(A \mid E_2) + P(E_3)P(A \mid E_3) \\ &= \frac{4}{9} \cdot \frac{3}{10} + \frac{2}{9} \cdot \frac{5}{10} + \frac{3}{9} \cdot \frac{8}{10} = \frac{23}{45} \end{aligned}$$

Example. 4·31. *The contents of urns I, II and III are as follows:*

1 white, 2 black and 3 red balls,

2 white, 1 black and 1 red balls, and

4 white, 5 black and 3 red balls.

One urn is chosen at random and two balls drawn. They happen to be white and red. What is the probability that they come from urns I, II or III ?

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

Solution. Let E_1, E_2, and E_3 denote the events that the urn I, II and III is chosen, respectively, and let A be the event that the two balls taken from the selected urn are white and red. Then

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$$

$$P(A \mid E_1) = \frac{1 \times 3}{{}^6C_2} = \frac{1}{5}, \quad P(A \mid E_2) = \frac{2 \times 1}{{}^4C_2} = \frac{1}{3},$$

and $$P(A \mid E_3) = \frac{4 \times 3}{{}^{12}C_2} = \frac{2}{11}$$

Hence

$$P(E_2 \mid A) = \frac{P(E_2)\, P(A \mid E_2)}{\sum_{i=1}^{3} P(E_i)\, P(A \mid E_i)}$$

$$= \frac{\frac{1}{3} \times \frac{1}{3}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11}} = \frac{55}{118}$$

Similarly

$$P(E_3 \mid A) = \frac{\frac{1}{3} \times \frac{2}{11}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{11}} = \frac{30}{118}$$

$$\therefore \quad P(E_1 \mid A) = 1 - \frac{55}{118} - \frac{30}{118} = \frac{33}{118}$$

Example. 4·32. *In answering a question on a multiple choice test a student either knows the answer or he guesses. Let p be the probability that he knows the answer and 1−p the probability that he guesses. Assume that a student who guesses at the answer will be correct with probability 1/5, where 5 is the number of multiple-choice alternatives. What is the conditional probability that a student knew the answer to a question given that he answered it correctly?*

[Delhi Univ. B.Sc. (Maths Hons.), 1985]

Solution. Let us define the following events:

E_1 : The student knew the right answer.

E_2 : The student guesses the right answer.

A : The student gets the right answer.

Then we are given

$P(E_1) = p, \quad P(E_2) = 1 - p, \quad P(A \mid E_2) = 1/5$

$P(A \mid E_1) = P$ [student gets the right answer given that he knew the right answer] $= 1$

We want $P(E_1 \mid A)$.

Using Bayes' rule, we get :

$$P(E_1 \mid A) = \frac{P(E_1)\, P(A \mid E_1)}{P(E_1)\, P(A \mid E_1) + P(E_2)\, P(A \mid E_2)} = \frac{p \times 1}{p \times 1 + (1-p) \times \frac{1}{5}} = \frac{5p}{4p+1}$$

Example 4·33. *In a bolt factory machines A, B and C manufacture respectively 25%, 35% and 40% of the total. Of their output 5, 4, 2 per cent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by machines A, B and C?*

Solution. Let E_1, E_2 and E_3 denote the events that a bolt selected at random is manufactured by the machines A, B and C respectively and let E denote the event of its being defective. Then we have

$$P(E_1) = 0{\cdot}25,\ P(E_2) = 0{\cdot}35,\ P(E_3) = 0{\cdot}40$$

The probability of drawing a defective bolt manufactured by machine A is $P(E \mid E_1) = 0{\cdot}05$.

Similarly, we have

$$P(E \mid E_2) = 0{\cdot}04, \text{ and } P(E \mid E_3) = 0{\cdot}02$$

Hence the probability that a defective bolt selected at random is manufactured by machine A is given by

$$P(E_1 \mid E) = \frac{P(E_1)\,P(E \mid E_1)}{\sum\limits_{i=1}^{3} P(E_i)\,P(E \mid E_i)}$$

$$= \frac{0{\cdot}25 \times 0{\cdot}05}{0{\cdot}25 \times 0{\cdot}05 + 0{\cdot}35 \times 0{\cdot}04 + 0{\cdot}40 \times 0{\cdot}02} = \frac{125}{345} = \frac{25}{69}$$

Similarly

$$P(E_2 \mid E) = \frac{0{\cdot}35 \times 0{\cdot}04}{0{\cdot}25 \times 0{\cdot}05 + 0{\cdot}35 \times 0{\cdot}04 + 0{\cdot}40 \times 0{\cdot}02} = \frac{140}{345} = \frac{28}{69}$$

and

$$P(E_3 \mid E) = 1 - [P(E_1 \mid E) + P(E_2 \mid E)] = 1 - \frac{25}{69} - \frac{28}{69} = \frac{16}{69}$$

This example illustrates one of the chief applications of Bayes Theorem.

EXERCISE 4 (d)

1. *(a)* State and prove Baye's Theorem.

(b) The set of events A_k, $(k = 1, 2, \ldots, n)$ are *(i)* exhaustive and *(ii)* pairwise mutually exclusive. If for all k the probabilities $P(A_k)$ and $P(E \mid A_k)$ are known, calculate $P(A_k \mid E)$, where E is an arbitrary event. Indicate where conditions *(i)* and *(ii)* are used.

(c) The events $E_1, E_2, \ldots, E_n$ are mutually exclusive and $E = E_1 \cup E_2 \cup \ldots \cup E_n$. Show that if $P(A \mid E_i) = P(B \mid E_i)$; $i = 1, 2, \ldots, n$, then $P(A \mid E) = P(B \mid E)$. Is this conclusion true if the events E_i are not mutually exclusive?

[Calcutta Univ. B.Sc. (Maths Hons.), 1990]

(d) What are the criticisms against the use of Bayes theorem in probability theory. **[Sri. Venketeswara Univ. B.Sc., 1991]**

(e) Using the fundamental addition and multiplication rules of probability, show that

$$P(B \mid A) = \frac{P(B)\,P(A \mid B)}{P(B)\,P(A \mid B) + P(\bar{B})\,P(A \mid \bar{B})}$$

where $\bar{B}$ is the event complementary to the event B.

[Delhi Univ. M.A. (Econ.), 1987]

2. *(a)* Two groups are competing for the positions on the Board of Directors of a corporation. The probabilities that the first and second groups will win are 0·6 and 0·4 respectively. Furthermore, if the first group wins the probability of introducing a new product is 0·8 and the corresponding probability if the second group wins is 0·3. What is the probability that the new product will be introduced?

Ans. $0{\cdot}6 \times 0{\cdot}8 + 0{\cdot}4 \times 0{\cdot}3 = 0{\cdot}6$

(b) The chances of X, Y, Z becoming managers of a certain company are 4:2:3. The probabilities that bonus scheme will be introduced if X, Y, Z become managers, are 0·3, 0·5 and 0·8 respectively. If the bonus scheme has been introduced, what is the probability that X is appointed as the manager.

Ans. 0·51

(c) A restaurant serves two special dishes, A and B to its customers consisting of 60% men and 40% women. 80% of men order dish A and the rest B. 70% of women order dish B and the rest A. In what ratio of A to B should the restaurant prepare the two dishes? **(Bangalore Univ. B.Sc., 1991)**

Ans. $P(A) = P[(A \cap M) \cup (A \cap W)] = 0{\cdot}6 \times 0{\cdot}8 + 0{\cdot}4 \times 0{\cdot}3 = 0{\cdot}6$

Similarly $P(B) = 0{\cdot}4$. Required ratio $= 0{\cdot}6 : 0{\cdot}4 = 3 : 2$.

3. *(a)* There are three urns having the following compositions of black and white balls.

Urn 1 : 7 white, 3 black balls
Urn 2 : 4 white, 6 black balls
Urn 3 : 2 white, 8 black balls.

One of these urns is chosen at random with probabilities 0·20, 0·60 and 0·20 respectively. From the chosen urn two balls are drawn at random without replacement. Calculate the probability that both these balls are white.

Ans. 8/45. **(Madurai Univ. B.Sc., 1991)**

(b) Bowl I contain 3 red chips and 7 blue chips, bowl II contain 6 red chips and 4 blue chips. A bowl is selected at random and then 1 chip is drawn from this bowl. *(i)* Compute the probability that this chip is red, *(ii)* Relative to the hypothesis that the chip is red, find the conditional probability that it is drawn from bowl II.

[Delhi Univ. B.Sc. (Maths Hons.) 1987]

(c) In a factory machines A and B are producing springs of the same type. Of this production, machines A and B produce 5% and 10% defective springs, respectively. Machines A and B produce 40% and 60% of the total output of the factory. One spring is selected at random and it is found to be defective. What is the possibility that this defective spring was produced by machine A?

[Delhi Univ. M.A. (Econ.), 1986]

(d) Urn A contains 2 white, 1 black and 3 red balls, urn B contains 3 white, 2 black and 4 red balls and urn C contains 4 white, 3 black and 2 red balls. One urn is chosen at random and 2 balls are drawn. They happen to be red and black. What

is the probability that both balls came from urn 'B' ?

[Madras U. B.Sc. April; 1989]

(e) Urn X_1, X_2, X_3, each contains 5 red and 3 white balls. Urns Y_1, Y_2, each contain 2 red and 4 white balls. An urn is selected at random and a ball is drawn. It is found to be red. Find the probability that the ball comes out of the urns of the first type. **[Bombay U. B.Sc., April 1992]**

(f) Two shipments of parts are received. The first shipment contains 1000 parts with 10% defectives and the second shipment contains 2000 parts with 5% defectives. One shipment is selected at random. Two parts are tested and found good. Find the probability (*a posterior*) that the tested parts were selected from the first shipment. **[Burdwan Univ. B.Sc. (Hons.), 1988]**

(g) There are three machines producing 10,000 ; 20,000 and 30,000 bullets per hour respectively. These machines are known to produce 5%, 4% and 2% defective bullets respectively. One bullet is taken at random from an hour's production of the three machines. What is the probability that it is defective? If the drawn bullet is defective, what is the probability that this was produced by the second machine? **[Delhi Univ. B.Sc. (Stat. Hons.), 1991]**

4. *(a)* Three urns are given each containing red and white chips as indicated.

Urn 1 : 6 red and 4 white.
Urn 2 : 2 red and 6 white.
Urn 3 : 1 red and 8 white.

(i) An urn is chosen at random and a ball is drawn from this urn. The ball is red. Find the probability that the urn chosen was urn I.

(ii) An urn is chosen at random and two balls are drawn without replacement from this urn. If both balls are red, find the probability that urn I was chosen. Under these conditions, what is the probability that urn III was chosen.

Ans. 108/173, 112/12, 0 **[Gauhati Univ. B.Sc., 1990]**

(b) There are ten urns of which each of three contains 1 white and 9 black balls, each of other three contains 9 white and 1 black ball, and of the remaining four each contains 5 white and 5 black balls. One of the urns is selected at random and a ball taken blindly from it turns out to be white. What is the probabililty that an urn containing 1 white and 9 black balls was selected? **(Agra Univ. B.Sc., 1991)**

Hint : $P(E_1) = \dfrac{3}{10}$, $P(E_2) = \dfrac{3}{10}$ and $P(E_3) = \dfrac{4}{10}$.

Let A be the event of drawing a white ball.

$$P(A) = \frac{3}{10} \times \frac{1}{10} + \frac{3}{10} \times \frac{9}{10} + \frac{4}{10} \times \frac{5}{10} = \frac{1}{2}$$

$$P(A \mid E_1) = \frac{1}{10} \text{ and } P(E_1 \mid A) = \frac{3}{50}$$

(c) It is known that an urn containing altogether 10 balls was filled in the following manner: A coin was tossed 10 times, and according as it showed heads or tails, one white or one black ball was put into the urn. Balls are drawn from this

urn one at a time, 10 times in succession (with replacement) and every one turns out to be white. Find the chance that the urn contains nothing but white balls.

Ans. 0·0702.

5. (*a*) From a vessel containing 3 white and 5 black balls, 4 balls are transferred into an empty vessel. From this vessel a ball is drawn and is found to be white. What is the probability that out of four balls transferred, 3 are white and 1 black. **[Delhi Uni. B.Sc. (Stat. Hons.), 1985]**

Hint. Let the five mutually exclusive events for the four balls transferred be E_0, E_1, E_2, E_3, and E_4, where E_i denotes the event that i white balls are transferred and let A be the event of drawing a white ball from the new vessel.

Then $$P(E_0) = \frac{^5C_4}{^8C_4},\ P(E_1) = \frac{^3C_1 \times {^5C_3}}{^8C_4},\ P(E_2) = \frac{^3C_2 \times {^5C_2}}{^8C_4}$$

$$P(E_3) = \frac{^3C_3 \times {^5C_1}}{^8C_4} \text{ and } P(E_4) = 0$$

Also $P(A \mid E_0) = 0,\ P(A \mid E_1) = \dfrac{1}{4},\ P(A \mid E_2) = \dfrac{2}{4},\ (A \mid E_3) = \dfrac{3}{4},$

and $P(A \mid E_4) = 1$. Hence $P(E_3 \mid A) = \dfrac{1}{7}$.

(*b*) The contents of the urns 1 and 2 are as follows :

Urn 1 : 4 white and 5 black balls.

Urn 2 : 3 white and 6 black balls.

One urn is chosen at random and a ball is drawn and its colour noted and replaced back to the urn. Again a ball is drawn from the same urn, colour noted and replaced. The process is repeated 4 times and as a result one ball of white colour and three balls of black colour are obtained. What is the probability that the urn chosen was the urn 1 ? **(Poona Univ. B.E., 1989)**

Hint. $P(E_1) = P(E_2) = 1/2,$

$P(A \mid E_1) = 4/9, \qquad 1 - P(A \mid E_1) = 5/9$

$P(A \mid E_2) = 1/3, \qquad 1 - P(A \mid E_2) = 2/3$

The probability that the urn chosen was the urn 1

$$= \frac{\frac{1}{2} \cdot \frac{4}{9} \left(\frac{5}{9}\right)^3}{\frac{1}{2} \cdot \frac{4}{9} \cdot \left(\frac{5}{9}\right)^3 + \frac{1}{2} \cdot \frac{1}{3} \cdot \left(\frac{2}{3}\right)^3}$$

(*c*) There are five urns numbered 1 to 5. Each urn contains 10 balls. The ith urn has i defective balls and $10 - i$ non-defective balls; $i = 1, 2, \ldots 5$. An urn is chosen at random and then a ball is selected at random from that urn. (*i*) What is the probability that a defective ball is selected ?

(*ii*) If the selected ball is defective, find the probability that it came from urn i, $(i = 1, 2, \ldots, 5)$. **[Delhi Univ. B.Sc. (Maths Hons.), 1987]**

Hint.: Define the following events :

E_i : ith urn is selected at random.

A : Defective ball is selected.

$P(E_i) = 1/5;\ i = 1, 2, ..., 5.$

$P(A \mid E_i) = P$ [Defective ball from ith urn] $= i/10, (i = 1, 2, ..., 5)$

$$P(E_i) \cdot P(A \mid E_i) = \frac{1}{5} \times \frac{i}{10} = \frac{i}{50}, \quad (i = 1, 2, ..., 5).$$

(i) $$P(A) = \sum_{i=1}^{5} P(E_i) P(A \mid E_i) = \sum_{i=1}^{5} \left(\frac{i}{50}\right) = \frac{1+2+3+4+5}{50} = \frac{3}{10}$$

(ii) $$P(E_i \mid A) = \frac{P(E_i) P(A \mid E_i)}{\sum_i P(E_i) P(A \mid E_i)} = \frac{i/50}{3/10} = \frac{i}{15}\ ;\ i = 1, 2, ..., 5.$$

For example, the probability that the defective ball came from 5th urn $= (5/15) = 1/3$.

6. *(a)* A bag contains six balls of different colours and a ball is drawn from it at random. A speaks truth thrice out of 4 times and B speaks truth 7 times out of 10 times. If both A and B say that a red ball was drawn, find the probability of their joint statement being true.

[Delhi Univ. B.Sc. (Stat. Hons.),1987; Kerala Univ. B.Sc.1988]

(b) A and B are two very weak students of Statistics and their chances of solving a problem correctly are 1/8 and 1/12 respectively. If the probability of their making a common mistake is 1/1001 and they obtain the same answer, find the chance that their answer is correct. **[Poona Univ. B.Sc., 1989]**

Ans. Reqd. Probability $= \dfrac{1/8 \times 1/12}{1/8 \times 1/12 + (1 - 1/8) \cdot (1 - 1/12) \cdot 1/1001} = \dfrac{13}{14}$

7. *(a)* Three boxes, practically indistinguishable in appearance, have two drawers each. Box I contains a gold coin in one and a silver coin in the other drawer, box II contains a gold coin in each drawer and box III contains a silver coin in each drawer. One box is chosen at random and one of its drawers is opened at random and a gold coin found. What is the probability that the other drawer contains a coin of silver? **(Gujarat Univ. B.Sc., 1992)**

Ans. 1/3, 1/3.

(b) Two cannons No. 1 and 2 fire at the same target. Cannon No. 1 gives on an average 9 shots in the time in which Cannon No. 2 fires 10 projectiles. But on an average 8 out of 10 projectiles from Cannon No. 1 and 7 out of 10 from Cannon No. 2 strike the target. In the course of shooting, the target is struck by one projectile. What is the probability of a projectile which has struck the target belonging to Cannon No. 2 ? **(Lucknow Univ. B.Sc., 1991)**

Ans. 0·493

(c) Suppose 5 men out of 100 and 25 women out of 10,000 are colour blind. A colour blind person is chosen at random. What is the probability of his being male? (Assume males and females to be in equal number.)

Hint. E_1 = Person is a male, E_2 = Person is a female.

A = Person is colour blind.

Then $P(E_1) = P(E_2) = ½$, $P(A \mid E_1) = 0{\cdot}05$, $P(A \mid E_2) = 0{\cdot}0025$.

Hence find $P(E_1 \mid A)$.

8. *(a)* Three machines X, Y, Z with capacities proportional to 2:3:4 are producting bullets. The probabilities that the machines produce defective are 0·1, 0·2 and 0·1 respectively. A bullet is taken from a day's production and found to be defective What is the probability that it came from machine X?

[Madras Univ. B.Sc., 1988]

(b) In a factory 2 machines M_1 and M_2 are used for manufacturing screws which may be uniquely classified as good or bad. M_1 produces per day n_1 boxes of screws, of which on the average, $p_1\%$ are bad while the corresponding numbers for M_2 are n_2 and p_2. From the total production of both M_1 and M_2 for a certain day, a box is chosen at random, a screw taken out of it and it is found to be bad. Find the chance that the selected box is manufactured *(i)* by M_1, *(ii)* M_2.

Ans. *(i)* $n_1 p_1/(n_1 p_1 + n_2 p_2)$, *(ii)* $n_2 p_2/(n_1 p_1 + n_2 p_2)$.

9. *(a)* A man is equally likley to choose any one of three routes A, B, C from his house to the railway station, and his choice of route is not influenced by the weather. If the weather is dry, the probabilities of missing the train by routes A, B, C are respectively 1/20, 1/10, 1/5. He sets out on a dry day and misses the train. What is the probability that the route chosen was C?

On a wet day, the respective probabilities of missing the train by routes A, B, C are 1/20, 1/5, 1/2 respectively. On the average, one day in four is wet. If he misses the train, what is the probability that the day was wet?

[Allahabad Univ. B.Sc., 1991]

(b) A doctor is to visit the patient and from past experience it is known that the probabilities that he will come by train, bus or scooter are respectively 3/10, 1/5, and 1/10, the probabililty that he will use some other means of transport being, therefore, 2/5. If he comes by train, the probability that he will be late is 1/4, if by bus 1/3 and if by scooter 1/12, if he uses some other means of transport it can be assumed that he will not be late. When he arrives he is late. What is the probability that *(i)* he comes by train *(ii)* he is not late?

[Burdwan Univ. B.Sc. (Hons.), 1990]

Ans. *(i)* 1/2, *(ii)* 9/34

10. State and prove Bayes rule and expalin why, in spite of its easy deductibility from the postulates of probability, it has been the subject of such extensive controversy.

In the chest X-ray tests, it is found that the probability of detection when a person has actually T.B. is 0·95 and probability of diagnosing incorrectly as having T.B. is 0·002. In a certain city 0·1% of the adult population is suspected to be suffering from T.B. If an adult is selected at random and is diagnosed as having

T.B. on the basis of the X-ray test, what is the probability of his actually having a T.B.? **(Nagpur Univ. B.E., 1991)**

Ans. 0·97

11. A certain transistor is manufactured at three factories at Barnsley, Bradford and Bristol. It is known that the Barnsley factory produces twice as many transistors as the Bradford one, which produces the same number as the Bristol one (during the same period). Experience also shows that 0·2% of the transistors produced at Barnsley and Bradford are faulty and so are 0·4% of those produced at Bristol.

A service engineer, while maintaining an electronic equipment, finds a defective transistor. What is the probability that the Bradford factory is to blame?

(Bangalore Univ. B.E., Oct. 1992)

12. The sample space consists of integers from 1 to $2n$ which are assigned probabilities proportional to their logarithms. Find the probabilities and show that the conditional probability of the integer 2, given that an even integer occurs, is

$$\frac{\log 2}{[n \log 2 + \log(n!)]}$$ **(Lucknow Univ. M.A., 1992)**

[**Hint.** Let E_i : the event that the integer $2i$ is drawn, $(i = 1, 2, 3, \ldots, n)$.

A : the event of drawing an even integer.

$$\Rightarrow \quad A = E_1 \cup E_2 \cup \ldots \cup E_n \quad \Rightarrow \quad P(A) = \sum_{i=1}^{n} P(E_i)$$

But $P(E_i) = k \log(2i)$ (Given)

$$\therefore \quad P(A) = k \sum_{i=1}^{n} \log(2i) = k \log \prod_{i=1}^{n} (2i) = k[n \log 2 + \log(n!)]$$

$$\therefore \quad P(E_i \mid A) = \frac{\log(2i)}{[n \log 2 + \log(n!)]}$$

13. In answering a question on a multiple choice test, an examinee either knows the answer (with probability p), or he guesses (with probability $1 - p$). Assume that the probability of answering a question correctly is unity for an examinee who knows the answer and $1/m$ for the examinee who guesses, where m is the number of multiple choice alternatives. Supposing an examinee answers a question correctly, what is the probability that he really knows the answer?

[Delhi Univ. M.C.A., 1990; M.Sc. (Stat.), 1989]

Hint. Let $E_1 =$ The examinee knows the answer,

$E_2 =$ The examinee guesses the answer,

and $A =$ The examinee answers correctly.

Then $P(E_1) = p$, $P(E_2) = 1 - p$, $P(A \mid E_1) = 1$ and $P(A \mid E_2) = 1/m$

Now use Bayes theorem to prove

$$P(E_1 \mid A) = \frac{mp}{1 + (m-1)p}$$

14. Die A has four red and two white faces whereas die B has two red and four white faces. A biased coin is flipped once. If it falls heads, the game continues by

throwing die A, if it falls tails die B is to be used.

(i) Show that the probability of getting a red face at any throw is 1/2.

(ii) If the first two throws resulted in red faces, what is the probability of red face at the 3rd throw?

(iii) If red face turns up at the first n throws, what is the probability that die A is being used?

Ans. *(ii)* 3/5 *(iii)* $\dfrac{2^n}{2^n+1}$

15. A manufacturing firm produces steel pipes in three plants with daily production volumes of 500, 1,000 and 2,000 units respectively. According to past experience it is known that the fraction of defective outputs produced by the three plants are respectively 0.005, 0.008 and 0.010. If a pipe is selected at random from a day's total production and found to be defective, from which plant does that pipe come?

Ans. Third plant.

16. A piece of mechanism consists of 11 components, 5 of type A, 3 of type B, 2 of type C and 1 of type D. The probability that any particular component will function for a period of 24 hours from the commencement of operations without breaking down is independent of whether or not any other component breaks down during that period and can be obtained from the following table:

Component type:	A	B	C	D
Probability:	0·6	0·7	0·3	0·2

(i) Calculate the probability that 2 components chosen at random from the 11 components will both function for a period of 24 hours from the commencement of operations without breaking down.

(ii) If at the end of 24 hours of operations neither of the 2 components chosen in *(i)* has broken down, what is the probability that they are both type C components.

Hint.

(i) Required probability
$$= \frac{1}{^{11}C_2}\,[\,^5C_2 \times (0{\cdot}6)^2 + {}^3C_2\,(0{\cdot}7)^2 + {}^2C_2\,(0{\cdot}3)^2$$
$$+ {}^5C_1 \times {}^3C_1 \times 0{\cdot}6 \times 0{\cdot}7 + {}^5C_1 \times {}^2C_1 \times (0{\cdot}6) \times (0{\cdot}3)$$
$$+ {}^5C_1 \times {}^1C_1 \times (0{\cdot}6) \times (0{\cdot}2) + {}^3C_1 \times {}^2C_1 \times 0{\cdot}7 \times 0{\cdot}3$$
$$+ {}^3C_1 \times {}^1C_1 \times 0{\cdot}7 \times 0{\cdot}2 + {}^2C_1 \times {}^1C_1 \times 0{\cdot}3 \times 0{\cdot}2]$$
$$= p \text{ (Say).}$$

(ii) Required probability (By Bayes theorem)
$$= \frac{^2C_2 \times (0{\cdot}3)^2}{p} = \frac{0{\cdot}09}{p}$$

4·10. Geometric probability. In remark 3, § 4·3·1 it was pointed out that the classical definition of probability fails if the total number of outcomes of an experiment is infinite. Thus, for example, if we are interested in finding the probability that a point selected at random in a given region will lie in a specified part of it, the classical definition of probability is modified and extended to what is called *geometrical probability or probability in continuum*. In this case, the general expression for probability 'p' is given by

$$p = \frac{\textit{Measure of specified part of the region}}{\textit{Measure of the whole region}}$$

where 'measure' refers to the length, area or volume of the region if we are dealing with one, two or three dimensional space respectively.

Example 4·34. *Two points are taken at random on the given straight line of length a. Prove that the probability of their distance exceeding a given length* c $(< a)$ is equal to $\left(1 - \frac{c}{a}\right)^2$.

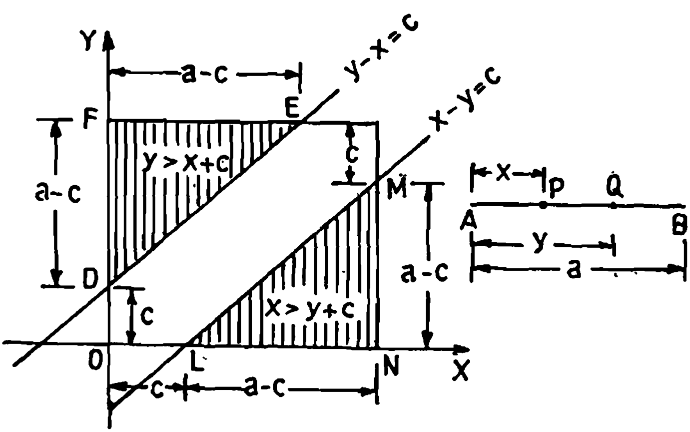
[Burdwan Univ. B.Sc. (Hons.), 1992; Delhi Univ. M.A. (Econ.), 1987]

Solution. Let P and Q be any two points taken at random on the given straight line AB of length 'a'. Let $AP = x$ and $AQ = y$,

$$(0 \le x \le a, 0 \le y \le a).$$

Then we want $P\{| x - y | > c\}$.

The probability can be easily calculated geometrically. Plotting the lines $x - y = c$ and $y - x = c$ along the co-ordinate axes, we get the following diagram:



Since $0 \le x \le a, 0 \le y \le a$, total area $= a \cdot a = a^2$.

Area favourable to the event $| x - y | > c$ is given by

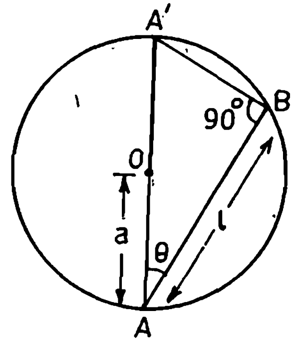
$$\Delta LMN + \Delta DEF = \frac{1}{2} LN \cdot MN + \frac{1}{2} EF \cdot DF$$

$$= \frac{1}{2}(a - c)^2 + \frac{1}{2}(a - c)^2 = (a - c)^2$$

$$P(| x - y | > c) = \frac{(a - c)^2}{a^2} = \left(1 - \frac{c}{a}\right)^2$$

Example 4·35. (Bertrand's Problem). *If a chord is taken at random in a circle, what is the chance that its length l is not less than 'a' , the radius of the circle?*

Solution. Let the chord AB make an angle θ with the diameter AOA' of the circle with centre O and radius $OA = a$. Obviously θ lies betwseen $-\pi/2$ and $\pi/2$. Since all the positions of the chord AB and consequently all the values of θ are equally likely, θ may be regarded as a random variable which is uniformly distributed *c.f.* § 8·1 over $(-\pi/2, \pi/2)$ with probability density function



$$f(\theta) = \frac{1}{\pi}\ ;\ -\pi/2 < \theta \leq \pi/2$$

$\angle ABA'$, being the angle in a semicircle, is a right angle. From $\Delta\, ABA'$ we have

$$\frac{AB}{AA'} = \cos\theta$$

$$\Rightarrow \qquad l = 2a\cos\theta$$

The required probability 'p' is given by

$$p = P(l \geq a) = P(2a\cos\theta \geq a)$$
$$= P(\cos\theta \geq 1/2) = P(|\theta| \leq \pi/3)$$
$$= \int_{-\pi/3}^{\pi/3} f(\theta)\, d\theta = \frac{1}{\pi}\int_{-\pi/3}^{\pi/3} d\theta = \frac{2}{3}$$

Example 4·36. *A rod of length 'a' is broken into three parts at random. What is the probability that a triangle can be formed from these parts?*

Solution. Let the lengths of the three parts of the rod be x, y and $a-(x+y)$. Obviously, we have

$$x > 0\ ;\ y > 0 \text{ and } x + y < a \quad \Rightarrow \quad y < a - x \qquad \ldots(*)$$

In order that these three parts form the sides of a triangle, we should have

$$\left.\begin{aligned} & x + y > a - (x+y) && \Rightarrow && y > \frac{a}{2} - x \\ \text{and} \quad & x + a - (x+y) > y && \Rightarrow && y < \frac{a}{2} \\ & y + a - (x+y) > x && \Rightarrow && y < \frac{a}{2} \end{aligned}\right\} \qquad \ldots(**)$$

since in a triangle, the sum of any two sides is greater than the third. Equivalently, (**) can be written as

$$\frac{a}{2} - x < y < \frac{a}{2} \quad \wedge \quad 0 < x < \frac{a}{2} \qquad \ldots(***)$$

Hence, on using (*) and (***), the required probability is given by

$$\frac{\int_0^{a/2}\int_{(a/2)-x}^{a/2} dy\,dx}{\int_0^{a}\int_0^{a-x} dy\,dx} = \frac{\int_0^{a/2}\left[\frac{a}{2}-\left(\frac{a}{2}-x\right)\right]dx}{\int_0^{a}(a-x)\,dx}$$

$$= \frac{\left|\frac{x^2}{2}\right|_0^{a/2}}{\left|\frac{-(a-x)^2}{2}\right|_0^{a}} = \frac{a^2/8}{a^2/2} = \frac{1}{4}$$

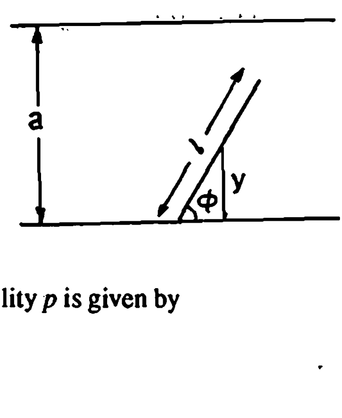
Example 4·37. (Buffon's Needle Problem). *A vertical board is ruled with horizontal parallel lines at constant distance 'a' apart. A needle of length l (< a) is thrown at random on the table. Find the probability that it will intersect one of the lines.*

Solution. Let y denote the distance from the centre of the needle to the nearest parallel and ϕ be angle formed by the needle with this parallel. The quantities y and ϕ fully determine the position of the needle. Obviously y ranges from 0 to $a/2$ (since $l < a$) and ϕ from 0 to π.

Since the needle is dropped randomly, all possible values of y and ϕ may be regarded as equally likely and consequently the joint probability density function $f(y, \phi)$ of y and ϕ is given by the uniform distribution (c.f. § 8.1) by

$$f(y, \phi) = k; \quad 0 \leq \phi \leq \pi, \quad 0 \leq y \leq a/2, \quad \ldots(*)$$

where k is a constant.



The needle will intersect one of the lines if the distance of its centre from the line is less than $\frac{1}{2} l \sin\phi$, *i.e.*, the required event can be represented by the inequality $0 < y < \frac{1}{2} l \sin\phi$. Hence the required probability p is given by

$$p = \frac{\int_0^{\pi}\int_0^{(l\sin\phi)/2} f(y, \phi)\,dy\,d\phi}{\int_0^{\pi}\int_0^{a/2} f(y, \phi)\,dy\,d\phi}$$

$$= \frac{\frac{l}{2}\int_0^{\pi}\sin\phi\,d\phi}{(a/2).\pi}$$

$$= \frac{l}{a\pi}\left|-\cos\phi\right|_0^{\pi} = \frac{2l}{a\pi}$$

EXERCISE 4 (e)

1. Two points are selected at random in a line AC of length 'a' so as to lie on the opposite sides of its mid-point O. Find the probability that the distance between them is less than $a/3$.

2. *(a)* Two points are selected at random on a line of length a. What is the probability that none of three sections in which the line is thus divided is less than $a/4$?

Ans. $1/16$.

(b) A rectilinear segment AB is divided by a point C into two parts $AC=a$, $CB=b$. Points X and Y are taken at random on AC and CB respectively. What is the probability that AX, XY and BY can form a triangle?

(c) ABG is a straight line such that AB is 6 inches and BG is 5 inches. A point Y is chosen at random on the BG part of the line. If C lies between B and G in such a way that $AC=t$ inches, find

(i) the probability that Y will lie in BC.

(ii) the probability that Y will lie in CG.

What can you say about the sum of these probabilities?

(d) The sides of a rectangle are taken at random each less than a and all lengths are equally likely. Find the chance that the diagonal is less than a.

3. *(a)* Three points are taken at random on the circumference of a circle. Find the chance that they lie on the same semi- circle.

(b) A chord is drawn at random in a given circle. What is the probability that it is greater than the side of an equilateral triangle inscribed in that circle?

(c) Show that the probability of choosing two points randomly from a line segment of length 2 inches and their being at a distance of at least 1 inch from each other is 1/4. [Delhi Univ. M.A. (Econ.), 1985]

4. A point is selected at random inside a circle. Find the probability that the point is closer to the centre of the circle than to its circumference.

5. One takes at random two points P and Q on a segment AB of length a

(i) What is the probability for the distance PQ being less than $b\ (<a)$?

(ii) Find the chance that the distance between them is greater than a given length b.

6. Two persons A and B, make an appointment to meet on a certain day at a certain place, but without fixing the time further than that it is to be between 2 p.m. and 3 p.m and that each is to wait not longer than ten minutes for the other. Assuming that each is independently equally likely to arrive at any time during the hour, find the probability that they meet.

Third person C, is to be at the same place from 2·10 p.m. until 2·40 p.m. on the same day. Find the probabilities of C being present when A and B are there together *(i)* When A and B remain after they meet, *(ii)* When A and B leave as soon as they meet.

Hint. Denote the times of arrival of A by x and of B by y. For the meeting to take place it is necessary and sufficient that

$$|x-y| < 10$$

We depict x and y as Cartesian coordinates in the plane; for the scale unit we take one minute. All possible outcomes can be described as points of a aquare with side 60. We shall finally get [*c.f.* Example 4·34, with $a = 60$, $c = 10$]

$$P[|x-y| < 10] = 1-(5/6)^2 = 11/36$$

7. The outcome of an experiment are represented by points in the square bounded by $x=0$, $x=2$ and $y=2$ in the xy-plane. If the probability is distributed uniformly, determine the probability that $x^2 + y^2 > 1$

Hint.

Required probability $P(E) = \int\limits_E \frac{1}{4}\, dx\, dy = 1 - \int\limits_{E'} \frac{1}{4}\, dx\, dy$

where E is the region for which $x^2+y^2 > 1$ and E' is the region for which $x^2+y^2 \le 1$.

$$\therefore \quad 4P(E) = 4 - \int_0^1 \int_0^1 dx\, dy = 3 \quad \Rightarrow \quad P(E) = \frac{3}{4}$$

8. A floor is paved with tiles, each tile being a parallelogram such that the distance between pairs of opposite sides are a and b respectively, the length of the diagonal being l. A stick of length c falls on the floor parallel to the diagonal. Show that the probability that it will lie entirely on one tile is

$$\left(1-\frac{c}{l}\right)^2$$

If a circle of diameter d is thrown on the floor, show that the probability that it will lie on one tile is

$$\left(1-\frac{d}{a}\right)\left(1-\frac{d}{b}\right)$$

9. Circular discs of radius r are thrown at random on to a plane circular table of radius R which is surrounded by a border of uniform width r lying in the same plane as the table. If the discs are thrown independently and at random, and N stay on the table, show that the probability that a fixed point on the table but not on the border, will be covered is

$$1-\left(1-\frac{r^2}{(R+r)^2}\right)^N$$

SOME MISCELLANEOUS EXAMPLES

Example 4·38. *A die is loaded in such a manner that for n=1, 2, 3, 4, 5, 6 the probability of the face marked n, landing on top when the die is rolled is proportional to n. Find the probability that an odd number will appear on tossing the die.*

[Madras Univ. B.Sc. (Stat. Main), 1987]

Solution. Here we are given

$P(n) \propto n$ or $P(n) = kn$, where k is the constant of proportionality.

Also $P(1) + P(2) + ... P(6) = 1 \Rightarrow k(1+2+3+4+5+6) = 1$ or $k = 1/21$

Required Probability $= P(1) + P(3) + P(5) = \dfrac{1+3+5}{21} = \dfrac{3}{7}$

Example 4·39. *In terms of probability :*

$p_1 = P(A)$, $p_2 = P(B)$, $p_3 = P(A \cap B)$, $(p_1, p_2, p_3 > 0)$

Express the following in terms of p_1, p_2, p_3

(a) $P(\overline{A \cup B})$, *(b)* $P(\bar{A} \cup \bar{B})$, *(c)* $P(\bar{A} \cap B)$, *(d)* $P(\bar{A} \cup B)$, *(e)* $P(\bar{A} \cap \bar{B})$

(f) $P(A \cap \bar{B})$, *(g)* $P(A \mid B)$, *(h)* $P(B \mid \bar{A})$, *(i)* $P[\bar{A} \cap (A \cup B)]$

Solution.

(a) $P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(AB)]$
$= 1 - p_1 - p_2 + p_3$.

(b) $P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - p_3$

(c) $P(\bar{A} \cap B) = P(B - AB) = P(B) - P(A \cap B) = p_2 - p_3$

(d) $P(\bar{A} \cup B) = P(\bar{A}) + P(B) - P(\bar{A} \cap B) = 1 - p_1 + p_2 - (p_2 - p_3)$
$= 1 - p_1 + p_3$

(e) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - p_1 - p_2 + p_3$. [Part *(a)*]

(f) $P(A \cap \bar{B}) = P(A - A \cap B) = P(A) - P(A \cap B) = p_1 - p_3$

(g) $P(A \mid B) = P(A \cap B)/P(B) = p_3/p_2$

(h) $P(B \mid \bar{A}) = P(\bar{A} \cap B)/P(\bar{A}) = (p_2 - p_3)/(1 - p_1)$

(i) $P[\bar{A} \cap (A \cup B)] = P[(\bar{A} \cap A) \cup (\bar{A} \cap B)]$

$= P(\bar{A} \cap B) = p_2 - p_3$ $[\because A \cap \bar{A} = \phi]$

Example 4·40. *Let $P(A) = p$, $P(A \mid B) = q$, $P(B \mid A) = r$. Find relations between the numbers p, q, r for the following cases :*

(a) Events A and B are mutually exclusive.

(b) A and B are mutually exclusive and collectively exhaustive.

(c) A is a subevent of B; B is a subevent of A.

(d) $\bar{A}$ and $\bar{B}$ are mutually exclusive.

[Delhi Univ. B.Sc. (Maths Hons.) 1985]

Solution. From given data : $P(A) = p$, $P(A \cap B) = P(A)\, P(B \mid A) = rp$

$$\therefore \quad P(B) = \frac{P(A \cap B)}{P(A \mid B)} = \frac{rp}{q}$$

(a) $P(A \cap B) = 0 \quad \Rightarrow \quad rp = 0.$

(b) $P(A \cap B) = 0$ and $P(A) + P(B) = 1$
$\Rightarrow \; p(q + r) = q;\; rp = 0 \;\Rightarrow\; pq = q \;\Rightarrow\; p = 1 \vee q = 0.$

(c) $A \subseteq B \Rightarrow A \cap B = A$ or $P(A \cap B) = P(A) \Rightarrow rp = p \Rightarrow r = 1 \vee p = 0.$
$B \subseteq A \Rightarrow A \cap B = B$ or $P(A \cap B) = P(B)$
$\Rightarrow \; rp = (rp/q)$ or $rp(q-1) = 0 \;\Rightarrow\; q = 1$

(d) $P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B) \;\Rightarrow\; 0 = 1 - [P(A) + P(B) - P(A \cap B)]$

So $P(A)+P(B)=1+P(A\cap B) \Rightarrow p[1+(r/q)]=1+rp$

$\therefore \quad p(q+r)=q(1+pr).$

Example 4·41. *(a) Twelve balls are distributed at random among three boxes. What is the probability that the first box will contain 3 balls?*

(b) If n biscuits be distributed among N persons, find the chance that a particular person receives r (< n) biscuits. **[Marathwada Univ. B.Sc. 1992]**

Solution. *(a)* Since each ball can go to any one of the three boxes, there are 3 ways in which a ball can go to any one of the three boxes. Hence there are 3^{12} ways in which 12 balls can be placed in the three boxes.

Number of ways in which 3 balls out of 12 can go to the first box is $^{12}C_3$. Now the remaining 9 balls are to be placed in 2 boxes and this can be done in 2^9 ways. Hence the total number of favourable cases $= {}^{12}C_3 \times 2^9$.

$\therefore$ Required probability $= \dfrac{^{12}C_3 \times 2^9}{3^{12}}$

(b) Take any one biscuit. This can be given to any one of the N beggars so that there are N ways of distributing any one biscuit. Hence the total number of ways in which n biscuit can be distributed at random among N beggars

$= N \,.\, N \ldots N \,(\, n \text{ times}) = N^n.$

'r' biscuits can be given to any particular beggar in ^nC_r ways. Now we are left with $(n-r)$ biscuits which are to be distributed among the remaining $(N-1)$ beggars and this can be done in $(N-1)^{n-r}$ ways.

$\therefore$ Number of favourable cases $= {}^nC_r \,.\, (N-1)^{n-r}$

Hence, required probability $= \dfrac{^nC_r\,(N-1)^{n-r}}{N^n}$

Example 4·42. *A car is parked among N cars in a row, not at either end. On his return the owner finds that exactly r of the N places are still occupied. What is the probability that both neighbouring places are empty?*

Solution. Since the owner finds on return that exactly r of the N places (including owner's car) are occupied, the exhaustive number of cases for such an arrangement is $^{N-1}C_{r-1}$ [since the remaining $r-1$ cars are to be parked in the remaining $N-1$ places and this can be done in $^{N-1}C_{r-1}$ ways].

Let A denote the event that both the neighbouring places to owner's car are empty. This requires the remaining $(r-1)$ cars to be parked in the remaining $N-3$ places and hence the number of cases favourable to A is $^{N-3}C_{r-1}$. Hence

$$P(A)=\frac{^{N-3}C_{r-1}}{^{N-1}C_{r-1}}=\frac{(N-r)(N-r-1)}{(N-1)(N-2)}$$

Example 4·43. *What is the probability that at least two out of n people have the same birthday? Assume 365 days in a year and that all days are equally likely.*

Solution. Since the birthday of any person can fall on any one of the 365 days, the exhaustive number of cases for the birthdays of n persons is 365^n.

If the birthdays of all the n persons fall on different days, then the number of favourable cases is

$$365\,(365-1)(365-2)\ldots[365-(n-1)],$$

because in this case the birthday of the first person can fall on any one of 365 days, the birthday of the second person can fall on any one of the remaining 364 days and so on.

Hence the probability (p) that birthdays of all the n persons are different is given by :

$$p = \frac{365\,(365-1)\,(365-2)\ldots[365-(n-1)]}{365^n}$$

$$= \left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\left(1-\frac{3}{365}\right)\ldots\left(1-\frac{n-1}{365}\right)$$

Hence the required probability that at least two persons have the same birthday is

$$1-p = 1-\left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\left(1-\frac{3}{365}\right)\ldots\left(1-\frac{n-1}{365}\right)$$

Example 4·44. *A five-figure number is formed by the digits 0, 1, 2, 3, 4 (without repetition). Find the probability that the number formed is divisible by 4.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

Solution. The total number of ways in which the five digits 0, 1, 2, 3, 4 can be arranged among themselves is 5!. Out of these, the number of arrangements which begin with 0 (and, therefore, will give only 4-digited numbers) is 4!. Hence the total number of five digited numbers that can be formed from the digits 0, 1, 2, 3, 4 is

$$5!-4! = 120-24 = 96$$

The number formed will be divisible by 4 if the number formed by the two digits on extreme right (*i.e.*, the digits in the unit and tens places) is divisible by 4. Such numbers are :

04 , 12 , 20 , 24 , 32 , and 40

If the numbers end in 04, the remaining three digits, *viz.*, 1, 2 and 3 can be arranged among themselves in 3 ! ways. Similarly, the number of arrangements of the numbers ending with 20 and 40 is 3 ! in each case.

If the numbers end with 12, the remaining three digits 0 , 3 , 4 can be arranged in 3 ! ways. Out of these we shall reject those numbers which start with 0 (*i.e.*, have 0 as the first digit). There are $(3-1)! = 2!$ such cases. Hence, the number of five digited numbers ending with 12 is

$$3!-2! = 6-2 = 4$$

Similarly the number of 5 digited numbers ending with 24 and 32 each is 4. Hence the total number of favourable cases is

$$3 \times 3\,! + 3 \times 4 = 18 + 12 = 30$$

Hence required probability $= \frac{30}{96} = \frac{5}{16}$

Example 4·45. *(Huyghen's problem). A and B throw alternately with a pair of ordinary dice. A wins if he throws 6 before B throws 7, and B wins if he throws 7 before A throws 6. If A begins, show that his chance of winning is 30 / 61*

[Delhi Univ. B.Sc. (Stat. Hons.), 1991; Delhi Univ. B.Sc., 1987]

Solution. Let E_1 denote the event of A's throwing '6' and E_2 the event of B's throwing '7' with a pair of dice. Then $\bar{E}_1$ and $\bar{E}_2$ are the complementary events.

'6' can be obtained with two dice in the following ways:

(1, 5), (5, 1), (2, 4), (4, 2), (3, 3), *i.e.*, in 5 distinct ways.

$$\therefore \quad P(E_1) = \frac{5}{36} \quad \text{and} \quad P(\bar{E}_1) = 1 - \frac{5}{36} = \frac{31}{36}$$

'7' can be obtained with two dice as follows:

(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3), *i.e.*, in 6 distinct ways.

$$\therefore \quad P(E_2) = \frac{6}{36} = \frac{1}{6} \quad \text{and} \quad P(\bar{E}_2) = 1 - \frac{1}{6} = \frac{5}{6}$$

If A starts the game, he will win in the following mutually exclusive ways:

(i) E_1 happens *(ii)* $\bar{E}_1 \cap \bar{E}_2 \cap E_1$ happens

(iii) $\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_1 \cap \bar{E}_2 \cap E_1$ happens, and so on.

Hence by addition theorem of probability, the required probability of A's winning, (say), $P(A)$ is given by

$$P(A) = P(i) + P(ii) + P(iii) + \ldots$$
$$= P(E_1) + P(\bar{E}_1 \cap \bar{E}_2 \cap E_1) + P(\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_1 \cap \bar{E}_2 \cap E_1) + \ldots$$
$$= P(E_1) + P(\bar{E}_1)\,P(\bar{E}_2)\,P(E_1) + P(\bar{E}_1)\,P(\bar{E}_2)\,P(\bar{E}_1)\,P(\bar{E}_2)\,P(E_1) + \ldots$$

(By compound probability theorem)

$$= \frac{5}{36} + \frac{31}{36} \times \frac{5}{6} \times \frac{5}{36} + \frac{31}{36} \times \frac{5}{6} \times \frac{31}{36} \times \frac{5}{6} \times \frac{5}{36} + \ldots$$

$$= \frac{5/36}{1 - \frac{31}{36} \times \frac{5}{6}} = \frac{30}{61}$$

Example 4·46. *A player tosses a coin and is to score one point for every head and two points for every tail turned up. He is to play on until his score reaches or passes n. If p_n is the chance of attaining exactly n score, show that*

$$p_n = \frac{1}{2}\,[\,p_{n-1} + p_{n-2}\,],$$

and hence find the value of p_n. **[Delhi Univ. B.Sc. (Stat. Hons.), 1992]**

Solution. The score n can be reached in the following two mutually exclusive ways:

(i) By throwing a tail when score is $(n-2)$, and

(ii) By throwing a head when score is $(n-1)$.

Hence by addition theorem of probability, we get

$$p_n = P(i) + P(ii) = \frac{1}{2} \cdot p_{n-2} + \frac{1}{2} \cdot p_{n-1} = \frac{1}{2}(p_{n-1} + p_{n-2}) \qquad \text{...}(*)$$

To find p_n explicitly, $(*)$ may be re-written as

$$\begin{aligned} p_n + \frac{1}{2} p_{n-1} &= p_{n-1} + \frac{1}{2} p_{n-2} \\ &= p_{n-2} + \frac{1}{2} p_{n-3} \\ &\dots \quad \dots \\ &\dots \quad \dots \\ &= p_2 + \frac{1}{2} p_1 \qquad \text{...}(**) \end{aligned}$$

Since the score 2 can be obtained as

(i) Head in first throw and head in 2nd throw,

(ii) Tail in the first throw, we have

$$p_2 = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \text{ and obviously } p_1 = \frac{1}{2}$$

Hence, from $(**)$, we get

$$p_n + \frac{1}{2} p_{n-1} = \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{2} = 1 = \frac{2}{3} + \frac{1}{3} = \frac{2}{3} + \frac{1}{2} \cdot \frac{2}{3}$$

$$\left.\begin{aligned} p_n - \tfrac{2}{3} &= (-\tfrac{1}{2})(p_{n-1} - \tfrac{2}{3}) \\ p_{n-1} - \tfrac{2}{3} &= (-\tfrac{1}{2})(p_{n-2} - \tfrac{2}{3}) \\ \vdots \qquad & \qquad \vdots \\ p_2 - \tfrac{2}{3} &= (-\tfrac{1}{2})(p_1 - \tfrac{2}{3}) \end{aligned}\right\}$$

Multiplying all the above equations, we get

$$\begin{aligned} p_n - \tfrac{2}{3} &= (-\tfrac{1}{2})^{n-1} (p_1 - \tfrac{2}{3}) \\ &= (-\tfrac{1}{2})^{n-1} (\tfrac{1}{2} - \tfrac{2}{3}) = (-1)^n \cdot \frac{1}{2^n} \cdot \frac{1}{3} \end{aligned}$$

$$\Rightarrow \qquad \begin{aligned} p_n &= \frac{2}{3} + (-1)^n \frac{1}{2^n} \cdot \frac{1}{3} \\ &= \frac{1}{3}\left[2 + (-1)^n \frac{1}{2^n}\right] \end{aligned}$$

Example 4·47. *A coin is tossed* $(m+n)$ *times,* (mn). *Show that the probability of at least* m *consecutive heads is* $\dfrac{n+2}{2^{m+1}}$.

[Kurukshetra Univ. M.Sc. 1990; Calcutta Univ. B.Sc.(Hons.), 1986]

Solution. Since $m > n$, only one sequence of m consecutive heads is possible. This sequence may start either with the first toss or second toss or third toss, and so on, the last one will be starting with $(n+1)th$ toss.

Let E_i denote the event that the sequence of m consecutive heads starts with ith toss. Then the required probability is

$$P(E_1) + P(E_2) + \dots + P(E_{n+1}) \qquad \dots(*)$$

Now $P(E_1) = P$ [Consecutive heads in first m tosses and head or tail in the rest]

$$= \left(\frac{1}{2}\right)^m$$

$P(E_2) = P$ [Tail in the first toss, followed by m consecutive heads and head or tail in the next]

$$= \frac{1}{2}\left(\frac{1}{2}\right)^m = \frac{1}{2^{m+1}}$$

In general,

$P(E_r) = P$ [tail in the $(r-1)$th trial followed by m consecutive heads and head or tail in the next]

$$= \frac{1}{2}\left(\frac{1}{2}\right)^m = \frac{1}{2^{m+1}}, \quad \forall \quad r = 2, 3, \dots, n+1.$$

Substituting in $(*)$,

$$\text{Required probability} = \frac{1}{2^m} + \frac{n}{2^{m+1}} = \frac{2+n}{2^{m+1}}$$

Example 4·48. *Cards are dealt one by one from a well-shuffled pack until an ace appears. Show that the probability that exactly n cards are dealt before the first ace appears is*

$$\frac{4(51-n)(50-n)(49-n)}{52.51.50.49}$$

[Delhi Univ. B.Sc. 1992]

Solution. Let E_i denote the event that an ace appears when the ith card is dealt. Then the required probability 'p' is given by

$p = P$ [Exactly n cards are dealt before the first ace appears]

$= P$ [The first ace appears at the $(n+1)$th dealing]

$$= P(\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_3 \cap \dots \cap \bar{E}_{n-1} \cap \bar{E}_n \cap E_{n+1})$$

$$= P(\bar{E}_1)\, P(\bar{E}_2 \mid \bar{E}_1)\, P(\bar{E}_3 \mid \bar{E}_1 \cap \bar{E}_2) \dots$$

$$\times P(\bar{E}_n \mid \bar{E}_1 \cap \bar{E}_2 \cap \dots \cap \bar{E}_{n-1}) \times P(E_{n+1} \mid \bar{E}_1 \cap \bar{E}_2 \cap \dots \cap \bar{E}_n) \qquad \dots(*)$$

Now

$$P(E_1) = \frac{4}{52} \quad \Rightarrow \quad P(\bar{E}_1) = \frac{48}{52}$$

$$P(E_2 \mid \bar{E}_1) = \frac{4}{51} \quad \Rightarrow \quad P(\bar{E}_2 \mid \bar{E}_1) = \frac{47}{51}$$

$$P(E_3 \mid \bar{E}_1 \cap \bar{E}_2) = \frac{4}{50} \quad \Rightarrow \quad P(\bar{E}_3 \mid \bar{E}_1 \cap \bar{E}_2) = \frac{46}{50}$$

$$\vdots \qquad\qquad \vdots$$

$$P(E_{n-1} \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-2}) = \frac{4}{52-(n-2)}$$

$$\therefore \quad P(\bar{E}_{n-1} \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-2}) = \frac{50-n}{52-(n-2)}$$

$$P(E_n \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-1}) = \frac{4}{52-(n-1)}$$

$$\therefore \quad P(\bar{E}_n \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-1}) = \frac{49-n}{52-(n-1)}$$

and $\quad P(E_{n+1} \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_n) = \dfrac{4}{52-n}$

Hence, from (*) we get

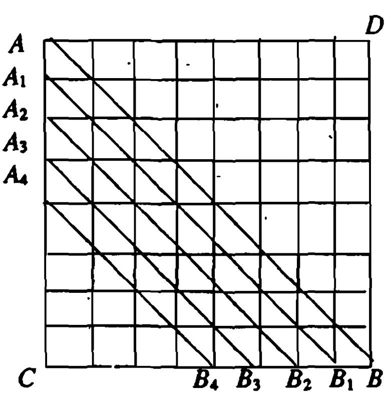
$$p = \left[\frac{48}{52} \times \frac{47}{51} \times \frac{46}{50} \times \frac{45}{49} \times \frac{44}{48} \times \frac{43}{47} \times \ldots \times \frac{52-n}{52-(n-4)} \right.$$
$$\left. \times \frac{51-n}{52-(n-3)} \times \frac{50-n}{52-(n-2)} \times \frac{49-n}{52-(n-1)} \times \frac{4}{52-n}\right]$$

$$= \frac{(51-n)(50-n)(49-n)\,4}{52 \times 51 \times 50 \times 49}$$

Example 4·49. *If four squares are chosen at random on a chess-board, find the chance that they should be in a diagonal line.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

Solution. In a chess-board there are $8 \times 8 = 64$ squares as shown in the following diagram.



Let us consider the number of ways in which the 4 squares selected at random are in a diagonal line parallel to AB. Consider the $\Delta\, ABC$. Number of ways in which 4 selected squares are along the lines A_4B_4, A_3B_3, A_2B_2, A_1B_1 and AB are ${}^4C_4$, ${}^5C_4$, ${}^6C_4$, ${}^7C_4$ and ${}^8C_4$ respectively.

Similarly, in $\Delta\, ABD$ there are an equal number of ways of selecting 4 squares in a diagonal line parallel to AB.

Hence, total number of ways in which the 4 selected squares are in a diagonal line parallel to AB are $2\,({}^4C_4 + {}^5C_4 + {}^6C_4 + {}^7C_4) + {}^8C_4$.

Since there is an equal number of ways in which 4 selected squares are in a diagonal line parallel to CD, the required number of favourable cases is given by

$$2\,[\,2(\,^4C_4 + {}^5C_4 + {}^6C_4 + {}^7C_4) + {}^8C_4\,]$$

Since 4 squares can be selected out of 64 in $^{64}C_4$ ways, the required probability is

$$= \frac{2\,[\,2(\,^4C_4 + {}^5C_4 + {}^6C_4 + {}^7C_4) + {}^8C_4]}{^{64}C_4}$$

$$= \frac{[4\,(\,1 + 5 + 15 + 35\,) + 140\,] \times 4\,!}{64 \times 63 \times 62 \times 61} = \frac{91}{158844}$$

Example 4·50. *An urn contains four tickets marked with numbers 112, 121, 211, 222 and one ticket is drawn at random. Let A_i, $(i=1, 2, 3)$ be the event that ith digit of the number of the ticket drawn is 1. Discuss the independence of the events A_1, A_2 and A_3.* **[Delhi Univ. B.Sc.(Stat. Hons.),1987; Poona Univ. B.Sc.,1986]**

Solution. We have

$$P(A_1) = \frac{2}{4} = \frac{1}{2} = P(A_2) = P(A_3)$$

$A_1 \cap A_2$ is the event that the first two digits in the number which the selected ticket bears are each equal to unity and the only favourable case is ticket with number 112.

$$\therefore \qquad P(A_1 \cap A_2) = \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2}$$
$$= P(A_1)\,P(A_2)$$

Similarly,

$$P(A_2 \cap A_3) = \frac{1}{4} = P(A_2)\,P(A_3)$$

and
$$P(A_3 \cap A_1) = \frac{1}{4} = P(A_3)\,P(A_1)$$

Thus we conclude that the events A_1, A_2 and A_3 are pairwise independent.

Now $P(A_1 \cap A_3 \cap A_3) = P$ {all the three digits in the number are 1's}

$$= P(\phi)$$
$$= 0 \neq P(A_1)\,P(A_2)\,P(A_3)$$

Hence A_1, A_2 and A_3 though pairwise independent are not mutually independent.

Example 4·51. *Two fair dice are thrown independently. Three events A, B and C are defined as follows:*

A : Odd face with first dice
B : Odd face with second dice
C : Sum of points on two dice is odd.

Are the events A, B and C mutually independent?

[Delhi Univ. B.Sc. (Stat. Hons.) 1983; M.S. Baroda Univ. B.Sc.1987]

Solution. Since each of the two dice can show any one of the six faces 1, 2, 3, 4, 5, 6, we get :

$$P(A) = \frac{3 \times 6}{36} = \frac{1}{2} \qquad [\because A = \{1, 3, 5\} \times \{1, 2, 3, 4, 5, 6\}]$$

$$P(B) = \frac{3 \times 6}{36} = \frac{1}{2} \qquad [\because B = \{1, 2, 3, 4, 5, 6\} \times \{1, 3, 5\}]$$

The sum of points on two dice will be odd if one shows odd number and the other shows even number. Hence favourable cases for C are :

(1, 2), (1, 4), (1, 6) ; (4, 1), (4, 3), (4, 5)
(2, 1), (2, 3), (2, 5) ; (5, 2), (5, 4), (5, 6)
(3, 2), (3, 4), (3, 6) ; (6, 1), (6, 3), (6, 5)

i.e., 18 cases in all.

Hence $P(C) = \frac{18}{36} = \frac{1}{2}$.

Cases favourable to the events $A \cap B, A \cap C, B \cap C$ and $A \cap B \cap C$ are given below :

Event	*Favourable cases*
$A \cap B$	(1,1), (1, 3), (1, 5), (3, 1), (3, 3), (3, 5), (5, 1) (5, 3) (5, 5), *i.e.,* 9 in all.
$A \cap C$	(1, 2), (1, 4), (1, 6), (3, 2), (3, 4), (3, 6), (5, 2), (5, 4) (5, 6), *i.e.,* 9 in all.
$B \cap C$	(2, 1), (4, 1), (6, 1) (2, 3), (4, 3), (6, 3), (2, 5), (4, 5), (6, 5), *i.e.,* 9 in all
$A \cap B \cap C$	Nil, because $A \cap B$ implies that sum of points on two dice is even and hence $(A \cap B) \cap C = \phi$

$$\therefore \qquad P(A \cap B) = \frac{9}{36} = \frac{1}{4} = P(A) \cdot P(B)$$

$$P(A \cap C) = \frac{9}{36} = \frac{1}{4} = P(A)\, P(C)$$

$$P(B \cap C) = \frac{9}{36} = \frac{1}{4} = P(B)\, P(C)$$

and $\qquad P(A \cap B \cap C) = P(\phi) = 0 \neq P(A)\, P(B)\, P(C)$

Hence the events A, B and C are pairwise independent but not mutually independent.

Example 4·52. *Let $A_1, A_2, \ldots, A_n$ be independent events and $P\{A_k\} = p_k$. Further, let p be the probability that none of the events occurs; then show that*

$$p \leq e^{-\Sigma p_k}$$

[Agra Univ. M.Sc., 1987]

Solution. We have

$$p = P(\bar{A}_1 \cap \bar{A}_2 \cap \ldots \cap \bar{A}_n)$$

$$= \prod_{i=1}^{n} P(\bar{A}_i) = \prod_{i=1}^{n} [1 - P(A_i)] = \prod_{i=1}^{n} (1 - p_i)$$

[since A_i 's are independent]

$$\leq \prod_{i=1}^{n} e^{-p_i} \qquad [\because 1 - x \leq e^{-x} \text{ for } 0 \leq x \leq 1 \text{ and } 0 \leq p_i \leq 1]$$

$$\Rightarrow \qquad p \leq \exp\left[-\sum_{i=1}^{n} p_i\right],$$

as desired.

Remark. We have

$$1 - x \leq e^{-x} \text{ for } 0 \leq x \leq 1 \qquad \ldots(*)$$

Proof. The inequality (*) is obvious for $x = 0$ and $x = 1$. Consider $0 < x < 1$. Then

$$\log(1 - x)^{-1} = -\log(1 - x)$$

$$= \left[x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \ldots\right], \qquad \ldots(**)$$

the expansion being valid since $0 < x < 1$. Further since $x > 0$, we get from (**)

$$\log(1 - x)^{-1} > x$$

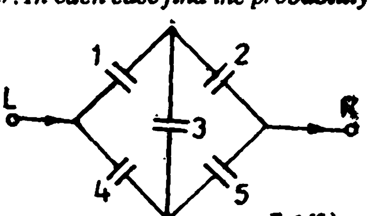
$$\Rightarrow \qquad -\log(1 - x) > x$$

$$\Rightarrow \qquad \log(1 - x) < -x$$

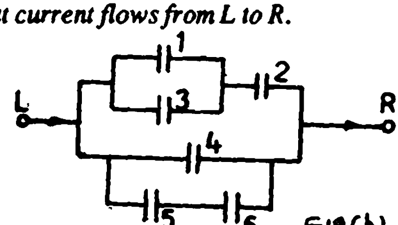
$$\Rightarrow \qquad 1 - x < e^{-x},$$

as desired.

Example 4·53. *In the following Fig.(a) and (b) assume that the probability of a relay being closed is P and that a relay is open or closed independently of any other. In each case find the probability that current flows from L to R.*



Fig(a)



Fig(b)

Solution. Let A_i denote the event that the relay i, ($i = 1, 2, \ldots, 6$) is closed. Let E be the event that current flows from L to R.

In Fig. (a) the current will flow from L to R if at least one of the circuits from L to R is closed. Thus for the current to flow from L to R we have the following favourable cases:

(i) $A_1 \cap A_2 = B_1$, (ii) $A_4 \cap A_5 = B_2$,

(iii) $A_1 \cap A_3 \cap A_5 = B_3$, (iv) $A_4 \cap A_3 \cap A_2 = B_4$,

The probability p_1 that current flows from L to R is given by

$$p_1 = P(B_1 \cup B_2 \cup B_3 \cup B_4) = \sum_i P(B_i) - \sum_{i<j} P(B_i \cap B_j) + \sum_{i<j<k} P(B_i \cap B_j \cap B_k) - P(B_1 \cap B_2 \cap B_3 \cap B_4) \quad ...(*)$$

Since the relays operate independently of each other, we have

$$P(B_1) = P(A_1 \cap A_2) = P(A_1) \,.\, P(A_2) = p \,.\, p = p^2$$
$$P(B_2) = P(A_4 \cap A_5) = P(A_4) \,.\, P(A_5) = p \,.\, p = p^2$$
$$P(B_3) = P(A_1)\, P(A_3)\, P(A_5) = p^3$$
$$P(B_4) = P(A_4)\, P(A_3)\, P(A_2) = p^3$$

Similarly

$$P(B_1 \cap B_2) = P(A_1 \cap A_2 \cap A_4 \cap A_5) = P(A_1)\, P(A_2)\, P(A_4)\, P(A_5) = p^4$$
$$P(B_1 \cap B_2 \cap B_3) = P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = p^5$$

and so on. Finally, substituting in (*), we get

$$p_1 = (p^2 + p^2 + p^3 + p^3) - (p^4 + p^4 + p^4 + p^4 + p^4 + p^5) + (p^5 + p^5 + p^5 + p^5) - p^5$$
$$= 2p^2 + 2p^3 - 5p^4 + 2p^5$$

In Fig. (b). Arguing as in the above case, the required probability p_2 that the current flows from L to R is given by

$$p_2 = P(E_1 \cup E_2 \cup E_3 \cup E_4)$$

where

$$E_1 = A_1 \cap A_2,\ E_2 = A_3 \cap A_2,\ E_3 = A_4,\ E_4 = A_5 \cap A_6$$

$$\therefore \quad p_2 = \sum_i P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - P(E_1 \cap E_2 \cap E_3 \cap E_4)$$
$$= (p^2 + p^2 + p + p^2) - (p^3 + p^3 + p^4 + p^3 + p^4 + p^3) + (p^4 + p^5 + p^5 + p^5) - p^6$$
$$= p + 3p^2 - 4p^3 - p^4 + 3p^5 - p^6$$

Matching Problem. Let us have n letters corresponding to which there exist n envelopes bearing different addresses. Considering various letters being put in various envelopes, *a match is said to occur* if a letter goes into the right envelope. (Alternatively, if in a party there are n persons with n different hats, *a match is said to occur* if in the process of selecting hats *at random*, the ith person rightly gets the ith hat.)

A match at the kth position for k=1, 2, ..., n. Let us first consider the event A_k when a match occurs at the kth place. For better understanding let us put the envelopes bearing numbners 1, 2, ..., n in ascending order. When A_k occurs, k th letter goes to the kth envelope but $(n-1)$ letters can go to the remaining $(n-1)$ envelopes in $(n-1)!$ ways.

$$\text{Hence} \quad P(A_k) = \frac{(n-1)!}{n!} = \frac{1}{n},$$

where $P(A_k)$ denotes the probability of the kth match. It is interesting to see that *$P(A_k)$ does not depend on k.*

Example 4·54. (a) *'n' different objects 1, 2, ..., n are distributed at random in n places marked 1, 2, ..., n. Find the probability that none of the objects occupies the place corresponding to its number.* **[Calcutta Univ. B.A.(Stat.Hons.)1986; Delhi Univ. B.Sc.(Maths Hons.), 1990; B.Sc.(Stat.Hons.) 1988]**

(b) If n letters are randomly placed in correctly addressed envelopes, prove that the probability that exactly r leters are placed in correct envelopes is given by

$$\frac{1}{r!} \sum_{k=0}^{n-r} (-1)^k \frac{1}{k!}\,; \; r = 1, 2, ..., n$$

[Bangalore Univ. B.Sc., 1987]

Solution (Probability of no match). Let E_i, $(i = 1, 2, ..., n)$ denote the event that the ith object occupies the place corresponding to its number so that $\overline{E}_i$, is the complementary event. Then the probability 'p' that none of the objects occupies the place corresponding to its number is given by

$$p = P(\overline{E}_1 \cap \overline{E}_2 \cap \overline{E}_3 \cap \dots \overline{E}_n)$$
$$= 1 - P\,\{\text{at least one of the objects occupies the place corresponding to its number}\}$$
$$= 1 - P(E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n)$$
$$= 1 - [\sum_{i=1}^{n} P(E_i) - \sum\sum_{\substack{i,j=1\\ i<j}}^{n} P(E_i \cap E_j) + \sum\sum\sum_{\substack{i,j,k=1\\ i<j<k}}^{n} P(E_i \cap E_j \cap E_k) - \dots$$
$$+ (-1)^{n-1} P(E_1 \cap E_2 \cap \dots \cap E_n)\,] \qquad \dots(*)$$

Now $\quad P(E_i) = \dfrac{1}{n}, \; \forall \; i$

$$P(E_i \cap E_j) = P(E_i)\,P(E_j \mid E_i)$$
$$= \frac{1}{n} \cdot \frac{1}{n-1}, \; \forall \;\; i, j \; (i<j)$$
$$P(E_i \cap E_j \cap E_k) = P(E_i)\,P(E_j \mid E_i)\,P(E_k \mid E_i \cap E_j)$$
$$= \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2}, \; \forall \;\; i, j, k \; (i<j<k)$$

and so on. Finally,

$$P(E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n) = \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} \dots \frac{1}{2} \cdot 1$$

Substituting in (*), we get

$$p = 1 - \left[{}^nC_1 \frac{1}{n} - {}^nC_2 \frac{1}{n(n-1)} + {}^nC_3 \frac{1}{n(n-1)(n-2)} - \dots + (-1)^{n-1} \frac{1}{n(n-1)\dots 3 \cdot 2 \cdot 1} \right]$$

$$= 1 - \left[1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n-1} \frac{1}{n!} \right]$$

$$= \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots + (-1)^n \frac{1}{n!}$$

$$= \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

Remark. For large n,

$$p = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots$$

$$= e^{-1} = 0{\cdot}36787$$

Hence the *probability of at least one match is*

$$1 - p = 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + \frac{(-1)^n}{n!}$$

$$= 1 - \frac{1}{e}, \text{ (for large } n)$$

(b) **[Probability of exactly r matches $\{r \leq (n-2)\}$]** Let A_i, $(i = 1, 2, \dots, n)$ denote the event that ith letter goes to the correct envelope. Then the probability that none of the n letters goes to the correct envelope is

$$P(\bar{A}_1 \cap \bar{A}_2 \cap \dots \cap \bar{A}_n) = \sum_{k=0}^{n} (-1)^k / k! \qquad \dots(**) \text{ [(c.f. part (a)]}$$

The probability that each of the 'r' letters is in the right envelope is $\frac{1}{n(n-1)(n-2)\dots(n-r+1)}$, and the probability that none of the remaining $(n-r)$ letters goes in the correct envelope is obtained by replacing n by $(n-r)$ in (**) and is thus given by $\sum_{k=0}^{n-r} \frac{(-1)^k}{k!}$. Hence by compound probability theorem, the probability that out of n letters exactly r letters go to correct envelopes, (in a specified order), is

$$\frac{1}{n(n-1)(n-2)\dots(n-r+1)} \sum_{k=0}^{n-r} \frac{(-1)^k}{k!}\ ; \quad r \leq n-2.$$

Since r letters can go to n envelopes in ${}^nC_r$ mutually exclusive ways, the required probability of exactly r letters going to correct envelopes, (in any order, whatsoever), is given by

$$^nC_r \times \frac{1}{n(n-1)(n-2)\dots(n-r+1)} \sum_{k=0}^{n-r} \frac{(-1)^k}{k!} = \frac{1}{r!} \sum_{k=0}^{n-r} (-1)^k \frac{1}{k!}$$

Example 4·55. *Each of the n urns contains 'a' white balls and 'b' black balls. One ball is transferred from the first urn to the second, then one ball from the latter into the third, and so on. If p_k is the probability of drawing a white ball from the kth urn, show that*

$$p_{k+1} = \frac{a+1}{a+b+1} p_k + \frac{a}{a+b+1} (1-p_k)$$

Hence for the last urn, prove that

$$p_n = \frac{a}{a+b}$$ **[Punjab Univ. B.Sc.(Maths Hons.),1988]**

Solution. The event of drawing a white ball from the kth urn can materialise in the following two ways:

(i) The ball transferred from the $(k-1)$th urn is white and then a white ball is drawn from the kth urn.

(ii) The ball transferred from the $(k-1)$th urn is black and then a white ball is drawn from the kth urn.

The probability of case *(i)* is $p_{k-1} \times \dfrac{a+1}{a+b+1}$,

since the probability of drawing a white ball from the $(k-1)$th urn is p_{k-1} and then the probability of drawing white ball from the kth urn is

$$\frac{a+1}{a+b+1}.$$

Since the probability of drawing a black ball from the $(k-1)$th urn is $[1-p_{k-1}]$ and then the probability of drawing a white ball from the kth urn is

$$\frac{a}{a+b+1},$$

the probability of case *(ii)* is given by

$$\frac{a}{a+b+1} [1-p_{k-1}]$$

Since the cases *(i)* and *(ii)* are mutually exclusive, we have by addition theorem of probability

$$p_k = \frac{a+1}{a+b+1} p_{k-1} + \frac{a}{a+b+1} [1-p_{k-1}] \quad \text{...(*)}$$

$$\therefore \quad p_k = \frac{1}{a+b+1} p_{k-1} + \frac{a}{a+b+1} \quad \text{...(1)}$$

Replacing k by $k+1$ in (*) we get the required result.

Changing k to $k-1, k-2, \dots$ and so on, we get

$$p_{k-1} = \frac{1}{a+b+1} p_{k-2} + \frac{a}{a+b+1} \quad \text{...(2)}$$

$$p_{k-2} = \frac{1}{a+b+1}\, p_{k-3} + \frac{a}{a+b+1} \qquad \text{...(3)}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$p_2 = \frac{1}{a+b+1}\, p_1 + \frac{a}{a+b+1} \qquad \text{...}(k-1)$$

But p_1 = Probability of drawing a white ball from the first urn $= \dfrac{a}{a+b}$.

Multiplying (1) by 1, (2) by $\dfrac{1}{a+b+1}$, (3) by $\left(\dfrac{1}{a+b+1}\right)^2$, ..., and $(k-1)$th equation by $\left(\dfrac{1}{a+b+1}\right)^{k-2}$ and adding, we get

$$p_k = \left(\frac{1}{a+b+1}\right)^{k-1} p_1 + \frac{a}{a+b+1}\left[1 + \frac{1}{a+b+1} + \frac{1}{(a+b+1)^2} + \dots + \left(\frac{1}{a+b+1}\right)^{k-2}\right]$$

$$= \left(\frac{1}{a+b+1}\right)^{k-1} \times \frac{a}{(a+b)} + \frac{a}{a+b+1}\left[\frac{1-\left(\dfrac{1}{a+b+1}\right)^{k-1}}{\left(1-\dfrac{1}{a+b+1}\right)}\right]$$

$$= \frac{a}{a+b}\left(\frac{1}{a+b+1}\right)^{k-1} + \frac{a}{a+b}\left[1-\left(\frac{1}{a+b+1}\right)^{k-1}\right]$$

$$= \frac{a}{a+b}\left[\left(\frac{1}{a+b+1}\right)^{k-1} + \left\{1-\left(\frac{1}{a+b+1}\right)^{k-1}\right\}\right]$$

$$= \frac{a}{a+b}, \quad (k = 1, 2, \dots, n)$$

Since the probability of drawing a white ball from the kth urn is independent of k, we have

$$p_n = \frac{a}{a+b}.$$

Example 4·56. *(i) Let the probability p_n that a family has exactly n children be αp^n when $n \geq 1$ and $p_0 = 1 - \alpha p(1 + p + p^2 + \dots)$. Suppose that all sex distributions of n children have the same probability. Show that for $k \geq 1$, the probability that a family contains exactly k boys is $2\alpha . p^k/(2-p)^{k+1}$.*

(ii) Given that a family includes at least one boy, show that the probability that there are two or more boys is $p/(2-p)$.

Solution. We are given

$$p_n = P\,[\text{that a family has exactly } n \text{ children}] = \alpha p^n,\ n \ge 1$$

and $\quad p_0 = 1 - \alpha p\,(1 + p + p^2 + \ldots)$

Let E_j be the event that the number of children in a family is j and let A be the event that a family contains exactly k boys. Then

$$P(E_j) = p_j;\ j = 0, 1, 2, \ldots$$

Now, since each child can have any of the two sex distributions (either boy or girl), the total number of possible distributions for a family to have 'j' children is 2^j.

$$\therefore \qquad P(A \mid E_j) = \frac{{}^jC_k}{2^j},\ j \ge k$$

and

$$\begin{aligned}
P(A) &= \sum_{j=k}^{\infty} P(E_j)\,P(A \mid E_j) = \sum_{j=k}^{\infty} p_j P(A \mid E_j)\\
&= \sum_{j=k}^{\infty} \alpha p^j \left[\frac{{}^jC_k}{2^j}\right], \quad j \ge k \ge 1\\
&= \alpha \sum_{j=k}^{\infty} \left(\frac{p}{2}\right)^j {}^jC_k\\
&= \alpha \sum_{r=0}^{\infty} {}^{k+r}C_k \left(\frac{p}{2}\right)^{k+r} && [\text{Put } j - k = r]\\
&= \alpha \left(\frac{p}{2}\right)^k \sum_{r=0}^{\infty} {}^{k+r}C_r \left(\frac{p}{2}\right)^r && [\because {}^nC_r = {}^nC_{n-r}]
\end{aligned}$$

We know that

$${}^{-n}C_r = (-1)^r \cdot {}^{n+r-1}C_r \quad \Rightarrow \quad (-1)^r \cdot {}^{-n}C_r = {}^{n+r-1}C_r$$

$$\therefore \qquad (-1)^r \cdot {}^{-(k+1)}C_r = {}^{k+r}C_r$$

Hence

$$\begin{aligned}
P(A) &= \alpha \left(\frac{p}{2}\right)^k \sum_{r=0}^{\infty} (-1)^r \cdot {}^{-(k+1)}C_r \left(\frac{p}{2}\right)^r\\
&= \alpha \left(\frac{p}{2}\right)^k \sum_{r=0}^{\infty} {}^{-(k+1)}C_r \left(-\frac{p}{2}\right)^r\\
&= \alpha \left(\frac{p}{2}\right)^k \left(1 - \frac{p}{2}\right)^{-(k+1)}\\
&= \alpha \left(\frac{p}{2}\right)^k \frac{2^{k+1}}{(2-p)^{k+1}} = \frac{2\alpha p^k}{(2-p)^{k+1}}
\end{aligned}$$

(b) Let B denote the event that a family includes at least one boy and C denote the event that a family has two or more boys. Then

$$P(B) = \sum_{k=1}^{\infty} P \text{ [family has exactly } k \text{ boys]}$$
$$= \sum_{k=1}^{\infty} \frac{2\alpha p^k}{(2-p)^{k+1}} = \frac{2\alpha}{2-p} \sum_{k=1}^{\infty} \left(\frac{p}{2-p}\right)^k$$
$$= \frac{2\alpha}{2-p} \times \frac{p/(2-p)}{1-[p/(2-p)]} = \frac{\alpha p}{(1-p)(2-p)}$$
$$P(C) = \sum_{k=2}^{\infty} P \text{ [family has exactly } k \text{ boys]}$$
$$= \sum_{k=2}^{\infty} \frac{2\alpha p^k}{(2-p)^{k+1}} = \frac{2\alpha}{2-p} \sum_{k=2}^{\infty} \left(\frac{p}{2-p}\right)^k$$
$$= \frac{2\alpha}{2-p} \cdot \frac{[p/(2-p)]^2}{1-[p/(2-p)]} = \frac{\alpha p^2}{(2-p)^2(1-p)}$$

Since $C \subset B$ and $B \cap C = C, P(B \cap C) = P(C) \Rightarrow P(B)P(C \mid B) = P(C)$

Therefore,

$$P(C \mid B) = \frac{P(C)}{P(B)} = \frac{\alpha p^2}{(2-p)^2(1-p)} \times \frac{(1-p)(2-p)}{\alpha p} = \frac{p}{2-p}$$

Example 4·57. *A slip of paper is given to person A who marks it either with a plus sign or a minus sign; the probability of his writing a plus sign is 1/3. A passes the slip to B, who may either leave it alone or change the sign before passing it to C. Next C passes the slip to D after perhaps changing the sign. Finally D passes it to a referee after perhaps changing the sign. The referee sees a plus sign on the slip. It is known that B, C and D each change the sign with probability 2/3. Find the probability that A originally wrote a plus.*

Solution. Let us define the following events :

E_1 : A wrote a plus sign; E_2 : A wrote a minus sign

E : The referee observes a plus sign on the slip.

We are given : $P(E_1) = 1/3$, $P(E_2) = 1 - 1/3 = 2/3$

We want $P(E_1 \mid E)$, which by Bayes rule is given by :

$$P(E_1 \mid E) = \frac{P(E_1)P(E \mid E_1)}{P(E_1)P(E \mid E_1) + P(E_2)P(E \mid E_2)} \quad ...(i)$$

$P(E \mid E_1) = P$ [Referee observes the plus sign given that 'A' wrote the plus sign on the slip]

$= P$ [(Plus sign was not changed at all) $\cup$ (Plus sign was changed exactly twice in passing from 'A' to referee through B, C and D)]

$= P(E_3 \cup E_4)$, (say).

$= P(E_3) + P(E_4)$ ...(ii)

Let A_1, A_2 and A_3 respectively denote the events that B, C and D change the sign on the slip. Then we are given

$P(A_1) = P(A_2) = P(A_3) = 2/3 \;;\; P(\bar{A}_1) = P(\bar{A}_2) = P(\bar{A}_3) = 1/3$

We have

$$P(E_3) = P(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3) = P(\bar{A}_1)\,P(\bar{A}_2)\,P(\bar{A}_3) = (1/3)^3 = 1/27$$

$$\begin{aligned} P(E_4) &= P[(A_1 A_2 \bar{A}_3) \cup (A_1 \bar{A}_2 A_3) \cup (\bar{A}_1 A_2 A_3)] \\ &= P(A_1 A_2 \bar{A}_3) + P(A_1 \bar{A}_2 A_3) + P(\bar{A}_1 A_2 A_3) \\ &= P(A_1)\,P(A_2)\,P(\bar{A}_3) + P(A_1)\,P(\bar{A}_2)\,P(A_3) + P(\bar{A}_1)\,P(A_2)\,P(A_3) \\ &= \frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{3} + \frac{2}{3}\cdot\frac{1}{3}\cdot\frac{2}{3} + \frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3} = \frac{4}{9} \end{aligned}$$

Substituting in (*ii*) we get

$$P(E \mid E_1) = \frac{1}{27} + \frac{4}{9} = \frac{13}{27} \qquad \text{...(iii)}$$

Similarly,

$$\begin{aligned} P(E \mid E_2) &= P\,[\text{Referee observes the plus sign given that 'A' wrote minus sign on the slip}] \\ &= P\,[(\text{Minus sign was changed exactly once}) \cup (\text{Minus sign was changed thrice})] \\ &= P(E_5 \cup E_6), \text{ (say)}, \\ &= P(E_5) + P(E_6) \qquad \text{...(iv)} \end{aligned}$$

$$\begin{aligned} P(E_5) &= P[(A_1 \bar{A}_2 \bar{A}_3) \cup (\bar{A}_1 A_2 \bar{A}_3) \cup (\bar{A}_1 \bar{A}_2 A_3)] \\ &= P(A_1)\,P(\bar{A}_2)\,P(\bar{A}_3) + P(\bar{A}_1)\,P(A_2)\,P(\bar{A}_3) + P(\bar{A}_1)\,P(\bar{A}_2)\,P(A_3) \\ &= \frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{2}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{1}{3}\cdot\frac{2}{3} = \frac{2}{9} \end{aligned}$$

$$P(E_6) = P(A_1 A_2 A_3) = P(A_1)\,P(A_2)\,P(A_3) = \frac{2}{3}\cdot\frac{2}{3}\cdot\frac{2}{3} = \frac{8}{27}$$

Substituting in (*iv*) we get :

$$P(E \mid E_2) = \frac{2}{9} + \frac{8}{27} = \frac{14}{27}. \qquad \text{...(v)}$$

Substituting from (*iii*) and (*v*) in (*i*) we get :

$$P(E_1 \mid E) = \frac{\frac{1}{3}\times\frac{13}{27}}{\frac{1}{3}\times\frac{13}{27} + \frac{2}{3}\times\frac{14}{27}} = \frac{13}{13 + 28} = \frac{13}{41}$$

Example 4·58. *Three urns of the same appearance have the following proportion of balls.*

First urn	:	*2 black*	*1 white*
Second Urn	:	*1 black*	*2 white*
Third urn	:	*2 black*	*2 white*

One of the urns is selected and one ball is drawn. It turns out to be white. What is the probability of drawing a white ball again, the first one not having been returned?

Solution. Let us define the events:

E_i = The event of selection of ith urn, $(i = 1, 2, 3)$

and A = The event of drawing a white ball.

Then

$$P(E_1) = P(E_2) = P(E_3) = 1/3$$

and $P(A \mid E_1) = 1/3$, $P(A \mid E_2) = 2/3$ and $P(A \mid E_3) = 1/2$

Let C denote the future event of drawing another white ball from the urns. Then

$$P(C \mid E_1 \cap A) = 0, P(C \mid E_2 \cap A) = \tfrac{1}{2}, \text{ and } P(C \mid E_3 \cap A) = \tfrac{1}{3}$$

$$\therefore \quad P(C \mid A) = \frac{\sum\limits_{i=1}^{3} P(E_i)\, P(A \mid E_i)\, P(C \mid E_i \cap A)}{\sum\limits_{i=1}^{3} P(E_i)\, P(A \mid E_i)}$$

$$= \frac{\frac{1}{3}\cdot\frac{1}{3}\cdot 0 + \frac{1}{3}\cdot\frac{2}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{2}{3} + \frac{1}{3}\cdot\frac{1}{2}} = \frac{1}{3}$$

MISCELLANEOUS EXERCISE ON CHAPTER IV

1. Probabilities of occurrence of n independent events $E_1, E_2, \ldots, E_n$ are $p_1, p_2, \ldots, p_n$ respectively. Find the probability of occurrence of the compound event in which $E_1, E_2, \ldots, E_r$ occur and $E_{r+1}, E_{r+2}, \ldots, E_n$ do not occur.

Ans. $\prod\limits_{i=1}^{r} p_i \times \prod\limits_{i=r+1}^{n} (1 - p_i)$

2. Prove that for any integer $m \geq 1$,

(a) $P\left(\bigcap\limits_{i=1}^{m} A_i\right) \leq P(A_i) \leq P\left(\bigcup\limits_{i=1}^{m} A_i\right) \leq \sum\limits_{i=1}^{m} P(A_i)$

(b) $P\left(\bigcap\limits_{i=1}^{m} A_i\right) \geq 1 - \sum\limits_{i=1}^{m} P(\overline{A}_i)$

3. Establish the inequalities :

$$P(A \cap B \cap C) \leq P(A \cap B) \leq P(A \cup B) \leq P(A \cup B \cup C) \leq P(A) + P(B) + P(C)$$

4. Let $A_1, A_2, \ldots, A_n$ be mutually independent events with $P(A_k) = p_k$, $k = 1, 2, \ldots, n$.

Let p be the probability that none of the events $A_1, A_2, \ldots, A_n$ occurs. Show that

$$p = \prod_{k=1}^{n} (1 - p_k) \leq \exp\left\{-\sum_{k=1}^{n} p_k\right\}$$

Use the above relation to compute the probability that in six tosses of a fair die, no "aces are obtained". Compare this with the upper bound given above. Show that if each p_k is small compared with n, the upper bound is a good approximation.

5. A and B play a match, the winner being the one who first wins two games in succession, no games being drawn. Their respective chances of winning a particular game are $p : q$. Find

(i) A's initial chance of winning.

(ii) A's chance of winning after having won the first game.

6. A carpenter has a tool chest with two compartments, each one having a lock. He has two keys for each lock, and he keeps all four keys in the same ring. His habitual procedure in opening a compartment is to select a key at random and try it. If it fails, he selects one of the remaining three and tries it and so on. Show that the probability that he succeeds on the first, second and third try is 1/2, 1/3, 1/6 respectively. **(Lucknow Univ. B.Sc., 1990)**

7. Three players A, B and C agree to play a series of games observing the following rules : two players participate in each game, while third is idle, and the game is to be won by one of them. The loser in each game quits and his place in the next game is taken by the player who was idle. The player who succeeds in winning over both of his opponents without interruption wins the whole series of games.

Supposing the probability for each player to win a single game is 1/2, and that the first game is played by A and B, find the probability for A, B and C respectively to win the whole series if the number of games is unlimited.

Ans. 5/14, 5/14, 2/7

8. In a certain group of mathematicians, 60 per cent have insufficient background of modern Algebra, 50 per cent have inadequate knowledge of Mathematical Statistics and 80 per cent are in either one or both of the two categories. What is the percentage of people who know Mathematical Statistics among those who have a sufficient background of Modern Algebra? **(Ans. 0·50)**

9. (a) If A has $(n+1)$ and B has n fair coins, which they flip, show that the probability that A gets more heads than B is $\frac{1}{2}$.

(b) A student is given a column of 10 dates and column of 10 events and is asked to match the correct date to each event. He is not allowed to use any item more than once. Consider the case where the student knows how to match four of the items but he is very doubtful of the remaining six. He decides to match these at random. Find the probabilities that he will correctly match (i) all the items, (ii) at least seven of the items, and (iii) at least five.

Ans. (a) $\frac{1}{6!}$, (b) $\frac{10}{6!}$, (c) $1-\frac{1}{6!}$

10. An astrologer claims that he can predict before birth the sex of a baby just to be born. Suppose that the astrologer has no real power but he tosses a coin just

once before every birth and if the head turns up he predicts a boy for that birth and if the tail turns up he predicts a girl. Let p be the probability of the event that at a certain birth a male child is born, and p' the probability of a head turning up in a single toss with astrologer's coin. Find the probability of a correct prediction and that of at least one correct prediction in n predictions.

11. From a pack of 52 cards an even number of cards is drawn. Show that the probability of half of these cards being red is

$$[52\,!/(26\,!)^2 - 1] / (2^{51} - 1)$$

12. A sportsman's chance of shooting an animal at a distance $r\ (> a)$ is a^2/r^2. He fires when $r = 2a$, and if he misses he reloads and fires when $r = 3a, 4a...$ If he misses at distance na, the animal escapes. Find the odds against the sportsman.

Ans. $n + 1 : n - 1$

Hint. P [Sportsman shoots at a distance ia] $= \dfrac{a^2}{(ia)^2} = \dfrac{1}{i^2}$

$\Rightarrow$ P [Sportsman misses the shot at a distance ia] $= 1 - \dfrac{1}{i^2}$

$$\therefore \quad P\,[\text{Animal escapes}] = \prod_{i=2}^{n}\left(1 - \frac{1}{i^2}\right) = \prod_{i=2}^{n}\left[\left(\frac{i-1}{i}\right)\left(\frac{i+1}{i}\right)\right]$$

$$= \prod_{i=2}^{n}\left(\frac{i-1}{i}\right)\prod_{i=2}^{n}\left(\frac{i+1}{i}\right) = \frac{n+1}{2n}$$

$$\text{Required ratio} = \frac{n+1}{2n} : \left(1 - \frac{n+1}{2n}\right) = (n+1) : (n-1)$$

13. *(a)* Pataudi, the captain of the Indian team, is reported to have observed the rule of calling 'heads' every time the toss was made during the five matches of the Test series with the Australian team. What is the probability of his winning the toss in all the five matches?

Ans. $(1/2)^5$

How will the probability be affected if

(i) he had made a rule of tossing a coin privately to decide whether to call "heads" or "tails" on each occasion.

(ii) the factors determining his choice were not predetermined but he called out whatever occurred to him on the spur of the moment?

(b) A lot contains 50 defective and 50 non-defective bulbs. Two bulbs are drawn at random one at a time, with replacement. The events A, B, C are defined as

$A =$ {The first bulb is defective}
$B =$ {The second bulb is non-defective}
$C =$ {The two bulbs are both defective or both non-defective}

Determine whether

(i) A, B, C are pairwise independent,

(ii) A, B, C are independent.

14. A, B and C are three urns which contain 2 white, 1 black, 3 white, 2 black and 2 white and 2 black balls, respectively. One ball is drawn from urn A and put into the urn B; then a ball is drawn from urn B and put into the urn C. Then a ball is drawn from urn C. Find the probability that the ball drawn is white.

Ans. 4/15.

15. An urn contains a white and b black balls and a series of drawings of one ball at a time is made, the ball removed being retrurned to the urn immediately after the next drawing is made. If p_n denotes the probability that the nth ball drawn is black, show that

$$p_n = (b - p_{n-1}) / (a + b - 1).$$

Hence find p_n.

16. A person is to be tested to see whether he can differentiate between the taste of two brands of cigarettes. If he cannot differentiate, it is assumed that the probability is one-half that he will identify a cigarette correctly. Under which of the following two procedures is there less chance that he will make all correct identifications when he actually cannot differentiate between the two brands?

(i) The subject is given four pairs each containing both brands of cigarettes (this is known to the subject), he must identify for each pair which cigarette represents each brand.

(ii) The subject is given eight cigarettes and is told that the first four are of one brand and the last four of the other brand.

How do you explain the difference in results despite the fact that eight cigarettes are tested in each case?

Ans. *(i)* 1/16 *(ii)* 1/2

17. (Sampling with replacement). A sample of size r is taken from a population of n people. Find the probability U_r that N given people will be included in the sample.

Ans. $U_r = \sum_{m=0}^{N} (-1)^m \binom{N}{m} \left(1 - \frac{m}{n}\right)^r$

18. In a lottery m tickest are drawn at a time out of the total number of n tickets, and returned before the next drawing is made. Show that the chance that in k drawings, each of the numbers 1, 2, 3, ..., n will appear at least once is given by

$$P_k = 1 - \binom{n}{1}\left(1 - \frac{m}{n}\right)^k + \binom{n}{2}\left(1 - \frac{m}{n}\right)^k \left(1 - \frac{m}{n-1}\right)^k - \ldots$$

[Nagpur Univ. M.Sc. 1987]

19. In a certain book of N pages, no page contains more than four errors, n_1 of them contain one error, n_2 contain two errors, n_3 contain three errors and n_4 contain four errors. Two copies of the book are opened at any two given pages. Show the probability that the number of errors in these two pages shall not exceed five is

$$1-\frac{1}{N^2}\ (n_3^2+n_4^2+2n_2 n_4+2n_3 n_4)$$

Hint. Let E_i I : the event that a page of first book contains i errors.

and E_i II : the event that a page of second book contains i errors.

P (No. of errors in the two pages shall not exceed 5)

$$= 1-P\,[E_2\,\mathrm{I}\,E_4\,\mathrm{II}+E_3\,\mathrm{I}\,E_4\,\mathrm{II}+E_4\,\mathrm{I}\,E_4\,\mathrm{II}$$
$$+E_3\,\mathrm{I}\,E_3\,\mathrm{II}+E_4\,\mathrm{I}\,E_3\,\mathrm{II}+E_4\,\mathrm{I}\,E_2\,\mathrm{II}\,]$$

20. *(a)* Of three independent events, the chance that the first only should happens is a, the chance of the second only is b and the chance of the third only is c. Show that the independent chances of the three events are respectively :

$$\frac{a}{a+x},\ \frac{b}{b+x},\ \frac{c}{c+x}$$

where x is the root of the equation

$$(a+x)(b+x)(c+x)=x^2$$

Hint. $P(E_1\cap\bar{E}_2\cap\bar{E}_3)=P(E_1)\,[1-P(E_2)]\,[1-P(E_3)]=a \quad ...(*)$

$P(\bar{E}_1\cap E_2\cap\bar{E}_3)=[1-P(E_1)]\,P(E_2)\,[1-P(E_3)]=b \quad ...(**)$

$P(\bar{E}_1\cap\bar{E}_2\cap E_3)=[1-P(E_1)]\,[1-P(E_2)]\,P(E_3)=c \quad ...(***)$

Multiplying (*), (**) and (***), we get

$$P(E_1)\,P(E_2)\,P(E_3)\,x^2 = abc,$$

where $x=[1-P(E_1)]\,[1-P(E_2)]\,[1-P(E_3)]$

Multiplying (*) by $[1-P(E_1)]$, we get

$$P(E_1)=\frac{a}{a+x}\text{, and so on.}$$

(b) Of three independent events, the probability that the first only should happens is 1/4, the probability that the second only should happen is 1/8, and the probability that the third only should happen is 1/12. Obtain the unconditional probabilities of the three events.

Ans. 1/2, 1/3, 1/4.

(c) A total of n shells are fired at a target. The probability of the ith shell hitting the target is p_i ; $i=1,2,3,...,n$. Assuming that the n firings are n mutually independent events, find the probability that at least two shells out of n hit the target. **[Calcutta Univ. B.Sc.(Maths Hons.), 1988]**

(d) An urn contains M balls numbered 1 to M, where the first K balls are defective and the remaining $M-K$ are non-defective. A sample of n balls is drawn from the urn. Let A_k be the event that the sample of n balls contains exactly k defectives. Find $P(A_k)$ when the sample is drawn *(i)* with replacement and, *(ii)* without replacement. **[Delhi Univ. B.Sc. (Maths Hons.), 1989]**

21. For three independent events A, B and C, the probability for A to occur is a, the probability that A, B and C will not occur is b, and the probability that at least one of the three events will not occur is c. If p denotes the probability that C occurs but neither A nor B occurs, prove that p satisfies the quadratic equation

$$ap^2 + [ab - (1-a)(a+c-1)]\,p + b(1-a)(1-c) = 0$$

and hence deduce that $c > \dfrac{(1-a)^2 + ab}{(1-a)}$

Further show that the probability of occurrence of C is $p/(p+b)$, and that of B's happening is $(1-c)(p+b)/ap$.

Hint. Let $P(A) = x$, $P(B) = y$ and $P(C) = z$

Then $\quad x = a,\ (1-x)(1-y)(1-z) = b,\ 1 - xyz = c$

and $\quad p = z(1-x)(1-y)$

Elimination of x, y and z gives quadratic equation in p.

22. *(a)* The chance of success in each trial is p. If p_k is the probability that there are even number of successes in k trials, prove that

$$p_k = p + p_{k-1}(1-2p)$$

Deduce that $\quad p_k = \frac{1}{2}[1 + (1-2p)^k]$

(b) If a day is dry, the conditional probability that the following day will also be dry is p; if a day is wet, the conditional probability that the following day will be dry is p'. If u_n is the probability that the nth day will be dry, prove that

$$u_n - (p - p')\,u_{n-1} - p' = 0\ ;\ n \geq 2$$

If the first day is dry, $p = 3/4$ and $p' = 1/4$, find u_n.

23. There are n similar biased dice such that the probability of obtaining a 6 with each one of them is the same and equal to p. If all the dice are rolled once, show that p_n, the probability that an odd number of 6's is obtained satisfies the difference equation

$$p_n + (2p-1)\,p_{n-1} = p$$

and hence derive an explicit expression for p_n.

Ans. $p_n = \frac{1}{2}[1 - (1-2p)^n]$

24. Suppose that each day the weather can be uniquely classified as 'fine' or 'bad'. Suppose further that the probability of having fine weather on the last day of a certain year is P_0 and we have the probability p that the weather on an arbitrary day will be of the same kind as on the preceding day. Let the probability of having fine weather on the nth day of the following year be P_n. Show that

$$P_n = (2p-1)\,P_{n-1} + (1-p)$$

Deduce that

$$P_3 = (2p-1)^3\left(P_0 - \frac{1}{2}\right) + \frac{1}{2}$$

25. A closet contains n pairs of shoes. If $2r$ shoes are chosen at random (with $2r < n$), what is the probability that there will be *(i)* no complete pair,

(ii) exactly one complete pair, *(iii)* exactly two complete pairs among them?

Hint. *(i)* $P(\text{no complete pair}) = \binom{n}{2r} 2^{2r} \div \binom{2n}{2r}$

(ii) $P(\text{exactly one complete pair}) = n \binom{n-1}{2r-2} 2^{2r-2} \div \binom{2n}{2r}$

and *(iii)* $P(\text{exactly two complete pairs}) = \binom{n}{2}\binom{n-2}{2r-4} 2^{2r-4} \div \binom{2n}{2r}$

26. Show that the probability of getting no right pair out of n, when the left foot shoes are paired randomly with the rigth foot shoes; is the sum of the first $(n+1)$ terms in the expansion of e^{-1}.

27. *(a)* In a town consisting of $(n+1)$ inhabitants, a person narrates a rumour to a second person, who in turn narrates it to a third person, and so on. At each step the recipient of the rumour is chosen at random from the n available persons, excluding the narrator himself. Find the probability that the rumour will be told r times without:

(i) returning to the originator,

(ii) being narrated to any person more than once.

(b) Do the above problem when, at each step the rumour is told by one person to a gathering of N randomly chosen people.

Ans. *(a)* *(i)* $\dfrac{n(n-1)^{r-1}}{n^r} = \left(1 - \dfrac{1}{n}\right)^{r-1}$; *(ii)* $\dfrac{n(n-1)(n-2)\ldots(n-r+1)}{n^r}$

(b) *(i)* $\left(1 - \dfrac{N}{n}\right)^{r-1}$; *(ii)* $\dfrac{\binom{n}{rN}}{\left[\binom{n}{N}\right]^r}$

28. What is the probability that *(i)* the birthdays of twelve people will fall in twelve different calendar months (assume equal probabilities for the twelve months) and *(ii)* the birthdays of six people will fall in exactly two calendar months?

Hint. *(i)* The birthday of the first person, for instance, can fall in 12 different ways and so for the second, and so on.

$\therefore$ The total number of cases $= 12^{12}$.

Now there are 12 months in which the birthday of one person can fall and 11 months in which the birthday of the second person can fall and 10 months for another third person, and so on.

$\therefore$ The total number of favourable cases $= 12.11.10\ldots3.2.1$

Hence the required probability $= \dfrac{12!}{12^{12}}$

(ii) The total number of ways in which the birthdays of 6 persons can fall in any of the month $= 12^6$.

$\therefore$ The required probability $= \dfrac{\binom{12}{2}(2^6 - 2)}{12^6}$

29. An elevator starts with 7 passengers and stops at 10 floors. What is the probability p that no two passengers leave at the same floor?

[Delhi Univ. M.C.A., 1988]

30. A bridge player knows that his two opponents have exactly five hearts between two of them. Each opponent has thirteen cards. What is the probability that there is three-two split on the hearts (that is one player has three hearts and the other two)? [Delhi Univ. B.Sc.(Maths Hons.), 1988]

31. An urn contains 2 white and 2 black balls. A ball is drawn at random. If it is white, it is not replaced into the urn. Otherwise it is replaced along with another ball of the same colour. The process is repeated. Find the probability that the third ball drawn is black. [Burdwan Univ. B.Sc. (Hons.), 1990]

Ans. $\dfrac{23}{30}$

32. There is a series of n urns. In the ith urn there are i white and $(n-i)$ black balls, $i = 1, 2, 3, ..., k$. One urn is chosen at random and 2 balls are drawn from it. Both turn out to be white. What is the probability that the jth urn was chosen, where j is a particular number between 3 and n.

Hint. Let E_j denote the event of selection of jth urn, $j = 3, 4, ..., n$ and A denote the event of drawing of 2 white balls, then

$$P(A \mid E_j) = \left(\frac{j}{n}\right)\left(\frac{j-1}{n-1}\right), \; P(E_j) = \frac{1}{n}, \; P(A) = \sum_{i=1}^{n} \frac{1}{n}\left(\frac{i}{n}\right)\left(\frac{i-1}{n-1}\right)$$

$$\therefore \qquad P(E_j \mid A) = \frac{\frac{1}{n}\left(\frac{j}{n}\right)\left(\frac{j-1}{n-1}\right)}{\sum\limits_{i=1}^{n}\left(\frac{1}{n}\right)\left(\frac{i}{n}\right)\left(\frac{i-1}{n-1}\right)}$$

33. There are $(N+1)$ identical urns marked $0, 1, 2, ..., N$ each of which contains N white and red balls. The kth urn contains k red and $N-k$ white balls, $(k = 0, 1, 2, ... N)$. An urn is chosen at random and n random drawings of a ball are made from it, the ball drawn being replaced after each draw. If the balls drawn are all red, show that the probability that the next drawing will also yield a red ball is approximately $(n+1)(n+2)$ when N is large.

34. A printing machine can print n letters, say $\alpha_1, \alpha_2, ..., \alpha_n$. It is operated by electrical impulses, each letter being produced by a different impulse. Assume that p is the constant probability of printing the correct letter and the impulses are independent. One of the n impulses, chosen at random, was fed into the machine twice and both times the letter α_1 was printed. Compute the probability that the impulse chosen was meant to print α_1. [Delhi Univ. M.Sc.(Stat.), 1981]

Ans. $(n-1)p^2/(np^2 - 2p + 1)$

35. Two players A and B agree to contest a match consisting of a series of games, the match to be won by the player who first wins three games, with the provision that if the players win two games each, the match is to continue until it

is won by one player winning two games more than his opponent. The probabililty of A winning any given game is p, and the games cannot be drawn.

(i) Prove that $f(p)$, the initial probability of A winning the match is given by:

$$f(p) = p^3 (4 - 5p + 2p^2)/(1 - 2p + 2p^2)$$

(ii) Show that the equation $f(p) = p$ has five real roots, of which three are admissible values of p. Find these three roots and explain their significance.

[Civil Services (Main), 1986]

36. Two players A and B start playing a series of games with Rs. a and b respectively. The stake is Re. 1 on a game and no game can be drawn. If the probability of A winning any game is a constant p, find the initial probability of his exhausting the funds of B or his own. Also show that if the resources of B are unlimited then

(i) A is certain to be ruined if $p = ½$, and

(ii) A has an even chance of escaping ruin if $p = 2^{1/a}/(1 + 2^{1/a})$.

Hint. Let u_n be the probability of A's final win when he has Rs. n.

Then $\qquad u_n = pu_{n+1} + (1-p)u_{n-1}$ where $u_0 = 0$ and $u_{a+b} = 1$

$$\therefore \qquad u_{n+1} - u_n = \left(\frac{1-p}{p}\right)(u_n - u_{n-1})$$

Hence $\quad u_{n+1} - u_n = \left(\frac{1-p}{p}\right)^n u_1$, by repeated application,

so that $$u_n = u_1\left[1 - \left(\frac{1-p}{p}\right)^n\right] \Big/ \left[1 - \left(\frac{1-p}{p}\right)\right]$$

Hence using $u_{a+b} = 1$, $$u_n = \left[1 - \left(\frac{1-p}{p}\right)^n\right] \Big/ \left[1 - \left(\frac{1-p}{p}\right)^{a+b}\right]$$

$\therefore$ Initial probability of A's win is $u_a = \dfrac{p^a - (1-p)^a}{p^{a+b} - (1-p)^{a+b}} \cdot p^b$

Probability of A's ruin $= 1 - u_a$.

For $\quad p = ½$, $u_a = \dfrac{a}{a+b} \to 0$ as $b \to \infty$ and for $p \neq ½$, $u_a = ½$ if $p = 2^{1/a}/(1 + 2^{1/a})$.

37. In a game of skill a player has probability 1/3, 5/12 and 1/4 of scoring 0, 1 and 2 points respectively at each trial, the game terminating on the first realization of a zero score at a trial. Assuming that the trials are independent, prove that the probability of the player obtaining a total score of n points is

$$u_n = \frac{3}{13}\left(\frac{3}{4}\right)^n + \frac{4}{39}\left(-\frac{1}{3}\right)^n$$

Hint. Event can materialize in the two mutually exclusive ways:

(i) at the $(n-1)$th trial, a score of $(n-1)$ points is obtained and a score of 1 point is obtained at the nth trial.

(ii) at the $(n-2)$th trial, a score of $(n-2)$ points is obtained and a score of 2 points is obtained at the last two trials.

Hence $u_n = \frac{5}{12} u_{n-1} + \frac{1}{4} u_{n-2}$ where $u_0 = \frac{1}{3}$, $u_1 = \frac{1}{3} \cdot \frac{5}{12} = \frac{5}{36}$

Also $u_n = \left(\frac{3}{4} - \frac{1}{3}\right) u_{n-1} + \frac{1}{4} u_{n-2} \Rightarrow u_n + \frac{1}{3} u_{n-1} = \frac{3}{4}\left(u_{n-1} + \frac{1}{3} u_{n-2}\right)$

This equation can be solved as a homogeneous difference equation of second order with the initial conditions

$$u_0 = \frac{1}{3},\ u_1 = \frac{1}{3} \cdot \frac{5}{12} = \frac{5}{36}$$

38. The following weather forecasting is used by an amateur forecaster. Each day is classified as 'dry' or 'wet' and the probability that any given day is same as the preceding one is assumed to be a constant p, $(0 < p < 1)$. Based on past records, it is supposed that January 1 has a probability β of being dry. Letting

β_n = Probability that nth day of the year is dry, obtain an expression for β_n in terms of β and p. Also evaluate $\lim\limits_{n \to \infty} \beta_n$.

Hint. $\beta_n = p \cdot \beta_{n-1} + (1-p)(1-\beta_{n-1})$

$\Rightarrow$ $\beta_n = (2p-1)\beta_{n-1} + (1-p)$; $n = 2, 3, 4, \ldots$

Ans. $\beta_n = (2p-1)^{n-1}(\beta - ½) + ½$; $\lim\limits_{n \to \infty} \beta_n = ½$

39. Two urns contain respectively 'a white and b black' and 'b white and a black' balls. A series of drawings is made according to the following rules:

(i) Each time only one ball is drawn and immediately returned to the same urn it came from.

(ii) If the ball drawn is white, the next drawing is made from the first urn.

(iii) If it is black, the next drawing is made from the second urn.

(iv) The first ball drawn comes from the first urn.

What is the probability that nth ball drawn will be white?

Hint. $p_r = P$ [Drawing a white ball at the rth draw].

$$p_r = \frac{a}{a+b} p_{r-1} + \frac{b}{a+b} \cdot \left(1 - p_{r-1}\right)$$

$$\Rightarrow \quad p_r = \frac{a-b}{a+b} \cdot p_{r-1} + \frac{b}{a+b}$$

Ans. $$p_n = \frac{1}{2} + \frac{1}{2}\left(\frac{a-b}{a+b}\right)^n$$

40. If a coin is tossed repeatedly, show that the probability of getting m heads before n tails is :

$$\frac{1}{2^{m+n-1}} \sum_{i=m}^{m+n-1} {}^{m+n-1}C_i.$$

[Burdwan Univ. (Maths Hons.), 1991]

OBJECTIVE TYPE QUESTIONS

I. Find out the correct answer from group Y for each item of group X:

Group X	Group Y
(a) At least one of the events A or B occurs.	(i) $(\overline{A} \cap B) \cup (A \cap \overline{B}) \cup (\overline{A} \cap \overline{B})$
(b) Neither A nor B occurs.	(ii) $(A \cup B) - (A \cap B)$
(c) Exactly one of the events A or B occurs.	(iii) $A \subset B$
(d) If event A occurs, so does B.	(iv) $B \not\subset A$
(e) Not more than one of the events A or B occur.	(v) $[A - (A \cap B)] \cup [B - (A \cap B)]$
	(vi) $A \cap \overline{B}$
	(vii) $1 - (A \cup \overline{B})$
	(viii) $A \cup B$
	(ix) $1 - (A \cup B)$

II. Match the correct expression of probabilities on the left:

(a) $P(\phi)$, where ϕ is null set	(i) $1 - P(A)$
(b) $P(A \mid B)\, P(B)$	(ii) $P(A \cap B)$
(c) $P(\overline{A})$	(iii) $P(A) - P(A \cap B)$
(d) $P(\overline{A} \cap \overline{B})$	(iv) 0
(e) $P(A \sim B)$	(v) $1 - P(A) - P(B) + P(A \cap B)$
	(vi) $P(A) + P(B) - P(A \cap B)$.

III. Given that A, B and C are mutually exclusive events, explain why the following are not permissible assignments of probabilities:

(i) $P(A) = 0{\cdot}24$, $P(B) = 0{\cdot}4$ and $P(A \cup C) = 0{\cdot}2$

(ii) $P(A) = 0{\cdot}4$, $P(B) = 0{\cdot}61$

(iii) $P(A) = 0{\cdot}6$, $P(A \cap \overline{B}) = 0{\cdot}5$

IV. In each of the following, indicate whether events A and B are: *(i)* independent, *(ii)* mutually exclusive, *(iii)* dependent but not mutually exclusive.

(a) $P(A \cap B) = 0$ (b) $P(A \cap B) = 0{\cdot}3$, $P(A) = 0{\cdot}45$

(c) $P(A \cup B) = 0{\cdot}85$, $P(A) = 0{\cdot}3$, $P(B) = 0{\cdot}6$

(d) $P(A \cup B) = 0{\cdot}70$, $P(A) = 0{\cdot}5$, $P(B) = 0{\cdot}4$

(e) $P(A \cup B) = 0{\cdot}90$, $P(A \mid B) = 0{\cdot}8$, $P(B) = 0{\cdot}5$.

V. Give the correct label as answer like a or b etc., for the following questions:

(i) The probability of drawing any one spade card from a pack of cards is

(a) $\frac{1}{52}$ (b) $\frac{1}{13}$ (c) $\frac{4}{13}$ (d) $\frac{1}{4}$

(ii) The probability of drawing one white ball from a bag containing 6 red, 8 black, 10 yellow and 1 green balls is

(a) $\frac{1}{25}$ (b) 0 (c) 1 (d) $\frac{24}{25}$ (e) $\frac{15}{20}$

(iii) A coin is tossed three times in succession, the number of sample points in sample space is

(a) 6 *(b)* 8 *(c)* 3

(iv) In the simultaneous tossing of two perfect coins, the probability of having at least one head is

(a) $\frac{1}{2}$ *(b)* $\frac{1}{4}$ *(c)* $\frac{3}{4}$ *(d)* 1

(v) In the simultaneous tossing of two perfect dice, the probability of obtaining 4 as the sum of the resultant faces is

(a) $\frac{4}{12}$ *(b)* $\frac{1}{12}$ *(c)* $\frac{3}{12}$ *(d)* $\frac{2}{12}$

(vi) A single letter is selected at random from the word 'probability'. The probability that it is a vowel is

(a) $\frac{3}{11}$ *(b)* $\frac{2}{11}$ *(c)* $\frac{4}{11}$ *(d)* 0

(vii) An urn contains 9 balls, two of which are red, three blue and four black. Three balls are drawn at random. The chance that they are of the same colour is

(a) $\frac{5}{84}$ *(b)* $\frac{3}{9}$ *(c)* $\frac{3}{7}$ *(d)* $\frac{7}{17}$

(viii) A number is chosen at random among the first 120 natural numbers. The probability of the number chosen being a multiple of 5 or 15 is

(a) $\frac{1}{5}$ *(b)* $\frac{1}{8}$ *(c)* $\frac{1}{16}$

(ix) If A and B are mutually exclusive events, then

(a) $P(A \cup B) = P(A) \cdot P(B)$

(b) $P(A \cup B) = P(A) + P(B)$, *(c)* $P(A \cup B) = 0$.

(x) If A and B are two independent events, the probability that both A and B occur is $\frac{1}{8}$ and the probability that neither of them occurs is $\frac{3}{8}$. The probability of the occurrence of A is :

(a) $\frac{1}{2}$, *(b)* $\frac{1}{3}$, *(c)* $\frac{1}{4}$, *(d)* $\frac{1}{5}$.

VI. Fill in the blanks :

(i) Two events are said to be equally likely if ......

(ii) A set of events is said to be independent if ......

(iii) If $P(A) \cdot P(B) \cdot P(C) = P(A \cap B \cap C)$, then the events A, B, C are ......

(iv) Two events A and B are mutually exclusive if $P(A \cap B) = \ldots$ and are independent if $P(A \cap B) = \ldots$

(v) The probability of getting a multiple of 2 in a throw of a dice is 1/2 and of getting a multiple of 3 is 1/3. Hence probability of getting a multiple of 2 or 3 is ......

(vi) Let A and B be independent events and suppose the event C has probability 0 or 1. Then A, B and C are ...... events.

(vii) If A, B, C are pairwise independent and A is independent of $B \cup C$, then A, B, C are ...... independent.

(viii) A man has tossed 2 fair dice. The conditional probability that he has tossed two sixes, given that he has tossed at least one six is ......

(ix) Let A and B be two events such that $P(A) = 0{\cdot}3$ and $P(A \cup B) = 0{\cdot}8$. If A and B are independent events then $P(B) = \ldots$

VII. Each of following statements is either true or false. If it is true prove it, otherwise, give a counter example to show that it is false.

(i) The probability of occurrence of at least one of two events is the sum of the probability of each of the two events.

(ii) Mutually exclusive events are independent.

(iii) For any two events A and B, $P(A \cap B)$ cannot be less than either $P(A)$ or $P(B)$.

(iv) The conditional probability of A given B is always greater than $P(A)$.

(v) If the occurrence of an event A implies the occurrence of another event B then $P(A)$ cannot exceed $P(B)$.

(vi) For any two events A and B, $P(A \cup B)$ cannot be greater then either $P(A)$ or $P(B)$.

(vii) Mutually exclusive events are not independent.

(viii) Pairwise independence does not necessarily imply mutual independence.

(ix) Let A and B be events neither of which has probability zero. Then if A and B are disjoint, A and B are independent.

(x) The probability of any event is always a proper fraction.

(xi) If $0 < P(B) < 1$ so that $P(A \mid B)$ and $P(A \mid \bar{B})$ are both defined, then $P(A) = P(B)\, P(A \mid B) + P(\bar{B})\, P(A \mid \bar{B})$.

(xii) For two events A and B if

$P(A) = P(A \mid B) = 1/4$ and $P(A \mid \bar{B}) = 1/2$, then

(a) A and B are mutually exclusive.

(b) A and B are independent.

(c) A is a sub-event of B.

(d) $P(\bar{A} \mid B) = 3/4$. **[Delhi Univ. B.Sc.(Stat. Hons.), 1992]**

(xiii) Two events can be independent and mutually exclusive simultaneously.

(xiv) Let A and B be events, neither of which has probability zero. Prove or disprove the following :

(a) If A and B are disjoint, A and B are independent.

(b) If A and B are independent, A and B are disjoint.

(xv) If $P(A) = 0$, then $A = \phi$.

Random Variables — Distribution Functions

5·1. Random Variable. Intuitively by a *random variable (r.v)* we mean a real number X connected with the outcome of a random experiment E. For example, if E consists of two tosses of a coin, we may consider the random variable which is the number of heads (0, 1 or 2).

Outcome :	*HH*	*HT*	*TH*	*TT*
Value of X :	*2*	*1*	*1*	*0*

Thus to each outcome ω, there corresponds a real number $X(\omega)$. Since the points of the sample space S correspond to outcomes, this means that a real number, which we denote by $X(\omega)$, is defined for each $\omega \in S$. From this standpoint, we define random variable to be a real function on S as follows:

" Let S be the sample space associated with a given random experiment. A real-valued function defined on S and taking values in R (– ∞ , ∞) is called a one-dimensional random variable. If the function values are ordered pairs of real numbers (i.e., vectors in two-space) the function is said to be a two- dimensional random variable. More generally, an n-dimensional random variable is simply a function whose domain is S and whose range is a collection of n-tuples of real numbers (vectors in n- space)."

For a mathematical and rigorous definition of the random variable, let us consider the probability space, the triplet (**S**, **B**, **P**), where S is the sample space, *viz.*, space of outcomes, **B** is the σ-field of subsets in S, and **P** is a probability function on **B**.

Def. A random variable (r.v.) is a function $X(\omega)$ with domain S and range $(-\infty, \infty)$ such that for every real number a, the event $[\omega : X(\omega) \leq a] \in \mathbf{B}$.

Remarks: 1. The refinement above is the same as saying that the function $X(\omega)$ is *measurable* real function on $(S, \mathbf{B})$.

2. We shall need to make probability statements about a random variable X such as $P\{X \leq a\}$. For the simple example given above we should write $P\{X \leq 1\} = P\{HH, HT, TH\} = 3/4$. That is, $P(X \leq a)$ is simply the probability of the set of outcomes ω for which $X(\omega) \leq a$ or

$$P(X \leq a) = P\{\omega : X(\omega) \leq a\}$$

Since **P** is a measure on $(S, \mathbf{B})$ *i.e.*, **P** is defined on subsets of **B**, the above probability will be defined *only* if $\{\omega : X(\omega) \leq a\} \in \mathbf{B}$, which implies that $X(\omega)$ is a measurable function on $(S, \mathbf{B})$.

3. One-dimensional random variables will be denoted by capital letters, $X, Y, Z, \ldots$ etc. A typical outcome of the experiment (*i.e.*, a typical element of the sample space) will be denoted by ω or e. Thus $X(\omega)$ represents the real number which the random variable X associates with the outcome ω. The values which $X, Y, Z, \ldots$ etc., can assume are denoted by lower case letters *viz.*, $x, y, z, \ldots$ etc.

4. *Notations*. If x is a real number, the set of all ω in S such that $X(\omega) = x$ is denoted briefly by writing $X = x$. Thus

$$P(X = x) = P\{\omega : X(\omega) = x\}$$

Similarly $\quad P(X \le a] = P\{\omega : X(\omega) \in [-\infty, a]\}$

and $\quad P[a < X \le b) = P(\omega : X(\omega) \in (a, b])$

Analogous meanings are given to

$$P(X = a \text{ or } X = b) = P\{(X = a) \cup (X = b)\},$$

$$P(X = a \text{ and } X = b) = P\{(X = a) \cap (X = b)\}, \quad \text{etc.}$$

Illustrations : 1. If a coin is tossed, then

$$S = \{\omega_1, \omega_2\} \text{ where } \omega_1 = H,\ \omega_2 = T$$

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = H \\ 0, & \text{if } \omega = T \end{cases}$$

$X(\omega)$ is a Bernoulli random variable. Here $X(\omega)$ takes only two values. A random variable which takes only a finite number of values is called *single*.

2. An experiment consists of rolling a die and reading the number of points on the upturned face. The most natural random variable X to consider is

$$X(\omega) = \omega\ ;\ \omega = 1, 2, \ldots, 6$$

If we are interested in whether the number of points is even or odd, we consider a random variable Y defined as follows :

$$Y(\omega) = \begin{cases} 0, & \text{if } \omega \text{ is even} \\ 1, & \text{if } \omega \text{ is odd} \end{cases}$$

3. If a dart is thrown at a circular target, the sample space S is the set of all points w on the target. By imagining a coordinate system placed on the target with the origin at the centre, we can assign various random variables to this experiment. A natural one is the two dimensional random variable which assigns to the point ω, its rectangular coordinates (x,y). Another is that which assigns ω its polar coordinates (r, θ). A one dimensional random variable assigns to each ω only one of the coordinates x or y (for cartesian system), r or θ (for polar system). The event **E**, "that the dart will land in the first quadrant" can be described by a random variable which assigns to each point w its polar coordinate θ so that $X(\omega) = \theta$ and then $\mathbf{E} = \{\omega : 0 \le X(\omega) \le \pi/2\}$.

4. If a pair of fair dice is tossed then $S = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$ and $n(S) = 36$. Let X be a random variable with image set

$$X(S) = \{1,2,3,4,5,6\}$$

$$P(X = 1) = P\{1,1\} = 1/36$$

$$P(X = 2) = P\{(2,1),(2,2),(1,2)\} = 3/36$$

$$P(X = 3) = P\{(3,1),(3,2),(3,3),(2,3),(1,3)\} = 5/36$$

$P(X=4) = P[(4,1),(4,2),(4,3),(4,4),(3,4),(2,4),(1,4)] = 7/36$

Similarly $P(X=5) = 9/36$ and $P(X=6) = 11/36$

Some theorems on Random Variables. Here we shall state (without proof) some of the fundamental results and theorems on random variables.

Theorem 5·1. A function $X(\omega)$ from S to $R(-\infty, \infty)$ is a random variable if and only if

$$\{\omega : X(\omega) < a\} \in \mathbf{B}$$

Theorem 5·2. If X_1 and X_2 are random variables and C is a constant then $CX_1, X_1 + X_2, X_1X_2$ are also random variables.

Remark. It will follow that $C_1X_1 + C_2X_2$ is a random variable for constants C_1 and C_2. In particular $X_1 - X_2$ is a r.v.

Theorem 5·3. If $\{X_n(\omega), n \geq 1\}$ are random variables then

$\sup\limits_{n} X_n(\omega)$, $\inf\limits_{n} X_n(\omega)$, $\limsup\limits_{n\to\infty} X_n(\omega)$ and $\liminf\limits_{n\to\infty} X_n(\omega)$ are all random variables, whenever they are finite for all ω.

Theorem 5·4. If X is a random variable then

(i) $\dfrac{1}{X}$ where $\left(\dfrac{1}{X}\right)(\omega) = \infty$ if $X(\omega) = 0$

(ii) $X_+(\omega) = \max[0, X(\omega)]$

(iii) $X_-(\omega) = -\min[0, X(\omega)]$

(iv) $|X|$

are random variables.

Theorem 5·5. If X_1 and X_2 are random variables then

(i) $\max[X_1, X_2]$ and (ii) $\min[X_1, X_2]$ are also random variables.

Theorem 5·6. If X is a r.v. and $f(\cdot)$ is a continuous function, then $f(X)$ is a r.v.

Theorem 5·7. If X is a r.v. and $f(\cdot)$ is an increasing function, then $f(X)$ *is a r.v.*

Corollary. If f is a function of bounded variations on every finite interval $[a,b]$, and X is a r.v. then $f(X)$ is a r.v.

(proofs of the above theorems are beyond the scope of this book)

EXERCISE 5 (a)

1. Let X be a one dimensional random variable. *(i)* If $a < b$, show that the two events $a < X \leq b$ and $X \leq a$ are disjoint, *(ii)* Determine the union of the two events in part *(i)*, *(iii)* show that $P(a < X \leq b) = P(X \leq b) - P(X \leq a)$.

2. Let a sample space S consist of three elements ω_1, ω_2, and ω_3. Let $P(\omega_1) = 1/4$, $P(\omega_2) = 1/2$ and $P(\omega_3) = 1/4$. If X is a random variable defined on S by $X(\omega_1) = 10$, $X(\omega_2) = -3$, $X(\omega_3) = 15$, find $P(-2 \leq X \leq 2)$.

3. Let $S = (e_1, e_2, ..., e_n)$ be the sample space of some experiment and let $E \subseteq S$ be some event associated with the experiment.

Define ψ_E, the *characteristic random variable* of E as follows :

$$\psi_E(e_i) = \begin{cases} 1 & \text{if } e_i \in E \\ 0 & \text{if } e_i \notin E. \end{cases}$$

In other words, ψ_E is equal to 1 if E occurs, and ψ_E is equal to 0 if E does not occur.

Verify the following properties of characteristic random variables :

(i) ψ_φ *is identically zero , i.e.,* $\psi_\varphi(e_i) = 0$; $i = 1, 2, \ldots, n$

(ii) ψ_S *is identically one , i.e.,* $\psi_S(e_i) = 1$; $i = 1, 2, \ldots, n$

(iii) $E = F \Rightarrow \psi_E(e_i) = \psi_F(e_i)$; $i = 1, 2, \ldots, n$ and *conversely*

(iv) If $E \subseteq F$ *then* $\psi_E(e_i) \le \psi_F(e_i)$; $i = 1, 2, \ldots, n$

(v) $\psi_E(e_i) + \psi_{\bar{E}}(e_i)$ *is identically* 1 : $i = 1, 2, ..., n$

(vi) $\psi_{E \cap F}(e_i) = \psi_E(e_i)\,\psi_F(e_i)$; $i = 1, 2, ..., n$

(vii) $\psi_{E \cup F}(e_i) = \psi_E(e_i) + \psi_F(e_i) - \psi_E(e_i)\,\psi_F(e_i)$, for $i = 1, 2, ..., n$.

5.2. Distribution Function. Let X be a r.v. on (S,B,P). Then the function :

$$F_X(x) = P(X \le x) = P\{\omega : X(\omega) \le x\}, \quad -\infty < x < \infty$$

is called the distribution function (d.f.) of X.

If clarity permits, we may write $F(x)$ instead of $F_X(x)$. ...(5·1)

5·2·1. Properties of Distribution Function. We now proceed to derive a number of properties common to all distribution functions.

Property 1. *If F is the d.f. of the r.v. X and if a < b, then*

$$P(a < X \le b) = F(b) - F(a)$$

Proof. The events '$a<X\le b$' and '$X\le a$' are disjoint and their union is the event '$X\le b$'. Hence by addition theorem of probability

$$P(a < X \le b) + P(X \le a) = P(X \le b)$$

$$\Rightarrow \quad P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a) \quad ...(5·2)$$

Cor. 1.

$$P(a \le X \le b) = P\{(X = a) \cup (a < X \le b)\}$$
$$= P(X = a) + P(a < X \le b)$$

(using additive property of P)

$$= P(X = a) + [F(b) - F(a)] \quad ...(5·2\,a)$$

Similarly, we get

$$P(a < X < b) = P(a < X \le b) - P(X = b)$$
$$= F(b) - F(a) - P(X = b) \quad ...(5·2\,b)$$

$$P(a \le X < b) = P(a < X < b) + P(X = a)$$

$$= F(b) - F(a) - P(X = b) + P(X = a) \quad ...(5·2\ c)$$

Remark. When $P(X = a) = 0$ and $P(X = b) = 0$, all four events $a \leq X \leq b$, $a < X < b$, $a \leq X < b$ and $a < X \leq b$ have the same probability $F(b) - F(a)$.

Property 2. *If F is the d.f. of one-dimensional r.v. X, then (i)* $0 \leq F(x) \leq 1$, *(ii)* $F(x) \leq F(y)$ *if* $x < y$.

In other words, all distribution functions are monotonically non-decreasing and lie between 0 and 1.

Proof. Using the axioms of certainty and non-negativity for the probability function **P**, part *(i)* follows triviality from the definition of $F(x)$.

For part *(ii)*, we have for $x < y$,

$$F(y) - F(x) = P(x < X \leq y) \geq 0 \qquad \text{(Property 1)}$$

$$\Rightarrow \qquad F(y) \geq F(x)$$

$$\Rightarrow \qquad F(x) \leq F(y) \text{ when } x < y \qquad ...(5·3)$$

Property 3. *If F is d.f. of one-dimensional r.v. X, then*

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

and

$$F(\infty) = \lim_{x \to \infty} F(x) = 1$$

Proof. Let us express the whole sample space S as a countable union of disjoint events as follows :

$$S = \left[\bigcup_{n=1}^{\infty} (-n < X \leq -n+1) \right] \cup \left[\bigcup_{n=0}^{\infty} (n < X \leq n+1) \right]$$

$$\Rightarrow P(S) = \sum_{n=1}^{\infty} P(-n < X \leq -n+1) + \sum_{n=0}^{\infty} P(n < X \leq n+1)$$

$$(\because \textbf{P} \text{ is additive})$$

$$\Rightarrow \qquad 1 = \lim_{a \to \infty} \sum_{n=1}^{a} [F(-n+1) - F(-n)]$$

$$+ \lim_{b \to \infty} \sum_{n=0}^{b} [F(n+1) - F(n)]$$

$$= \lim_{a \to \infty} [F(0) - F(-a)] + \lim_{b \to \infty} [F(b+1) - F(0)]$$

$$= [F(0) - F(-\infty)] + [F(\infty) - F(0)]$$

$$\therefore \qquad 1 = F(\infty) - F(-\infty) \qquad ...(*)$$

Since $-\infty < \infty$, $F(-\infty) \leq F(\infty)$. Also

$$F(-\infty) \geq 0 \text{ and } F(\infty) \leq 1 \qquad \text{(Property 2)}$$

$$\therefore \qquad 0 \leq F(-\infty) \leq F(\infty) \leq 1 \qquad (**)$$

(*) and (**) give $F(-\infty) = 0$ and $F(\infty) = 1$.

Remarks. 1. Discontinuities of $F(x)$ are at most countable.

2. $$F(a) - F(a-0) = \lim_{h \to 0} P(a-h \leq X \leq a),\ h > 0$$

$$\therefore \qquad F(a) - F(a-0) = P(X = a)$$

and $$F(a+0) - F(a) = \lim_{h \to 0} P(a \leq X \leq a+h) = 0,\ h > 0$$

$$\Rightarrow \qquad F(a+0) = F(a)$$

5·3. Discrete Random Variable. If a random variable takes at most a countable number of values, it is called a discrete random variable. *In other words, a real valued function defined on a discrete sample space is called a discrete random variable.*

5·3·1. Probability Mass Function (*and probability distribution of a discrete random variable*).

Suppose X is a one-dimensional discrete random variable taking at most a countably infinite number of values $x_1, x_2, \ldots$ With each possible outcome x_i, we associate a number $p_i = P(X = x_i) = p(x_i)$, called the probability of x_i. The numbers $p(x_i)$; $i = 1, 2, \ldots$ must satisfy the following conditions:

$$(i) \quad p(x_i) \geq 0 \ \forall \ i, \quad (ii) \quad \sum_{i=1}^{\infty} p(x_i) = 1$$

This function p is called the probability mass function of the random variable X and the set $\{x_i, p(x_i)\}$ is called the probability distribution (p.d.) of the r.v. X.

Remarks: 1. The set of values which X takes is called the *spectrum* of the random variable.

2. For discrete random variable, a knowledge of the probability mass function enables us to compute probabilities of arbitrary events. In fact, if E is a set of real numbers, we have

$$P(X \in E) = \sum_{x \in E \cap S} p(x), \text{ where } S \text{ is the sample space.}$$

Illustration. Toss of coin, $S = \{H, T\}$. Let X be the random variable defined by

$$X(H) = 1, \text{ i.e., } X = 1, \text{ if 'Head' occurs.}$$
$$X(T) = 0, \text{ i.e., } X = 0, \text{ if 'Tail' occurs.}$$

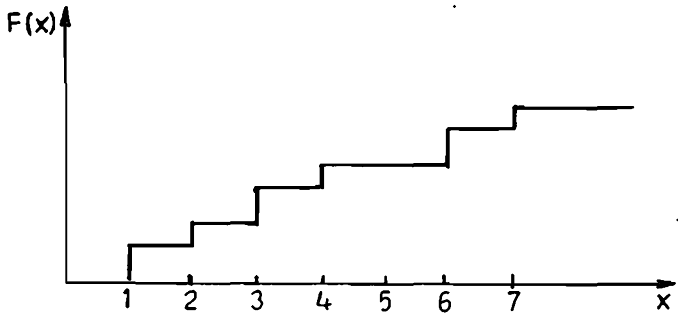
If the coin is 'fair' the probability function is given by

$$P(\{H\}) = P(\{T\}) = \tfrac{1}{2}$$

and we can speak of the probability distribution of the random variable X as

$$P(X = 1) = P(\{H\}) = \tfrac{1}{2},$$
$$P(X = 0) = P(\{T\}) = \tfrac{1}{2},$$

5·3·2. Discrete Distribution Function. In this case there are a countable number of points $x_1, x_2, x_3, \ldots$ and numbers $p_i \geq 0$, $\sum_{1}^{\infty} p_i = 1$ such that $F(X) = \sum_{(i : x_i \leq x)} p_i$. For example if x_i is just the integer i, $F(x)$ is a "*step function*" having jump p_i at i, and being constant between each pair of integers.



Theorem 5·5. $p(x_j) = P(X = x_j) = F(x_j) - F(x_{j-1})$, where F is the *d.f.* of X.

Proof. Let $x_1 < x_2 < \ldots$ We have

$$F(x_j) = P(X \leq x_j)$$

$$= \sum_{i=1}^{j} P(X = x_i) = \sum_{i=1}^{j} p(x_i)$$

and
$$F(x_{j-1}) = P(X \leq x_{j-1}) = \sum_{i=1}^{j-1} p(x_i)$$

$$\therefore \qquad F(x_j) - F(x_{j-1}) = p(x_j) \qquad \ldots(5·5)$$

Thus, given the distribution function of discrete random variable, we can compute its probability mass function.

Example 5·1. *An experiment consists of three independent tosses of a fair coin. Let*

X = The number of heads

Y = The number of head runs,

Z = The lenght of head runs,

a head run being defined as consecutive occurrence of at least two heads, its length then being the number of heads occurring together in three tosses of the coin.

Find the probability function of (i) X, (ii) Y, (iii) Z, (iv) $X+Y$ and (v) XY and construct probability tables and draw their probability charts.

Solution.

Table 1

S. No.	Elementary event	Random Variables				
		X	Y	Z	$X+Y$	XY
1	HHH	3	1	3	4	3
2	HHT	2	1	2	3	2
3	HTH	2	0	0	2	0
4	HTT	1	0	0	1	0
5	THH	2	1	2	3	2
6	THT	1	0	0	1	0
7	TTH	1	0	0	1	0
8	TTT	0	0	0	0	0

Here sample space is.

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

(i) Obviously X is a r.v. which can take the values 0, 1, 2, and 3

$$p(3) = P(HHH) = (1/2)^3 = 1/8$$

$$p(2) = P[HHT \cup HTH \cup THH]$$

$$= P(HHT) + P(HTH) + P(THH) = 1/8 + 1/8 + 1/8 = 3/8$$

Similarly $p(1) = 3/8$ and $p(0) = 1/8$.

These probabilities could also be obtained directly from the above table 1.

Table 2

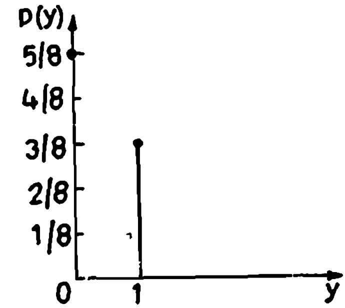
Probability table of X

Values of X (x)	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

Table 3

***(ii)* Probability Table of Y**

Values of Y, (y)	0 1
$p(y)$	5/8 3/8



Probability chart of Y

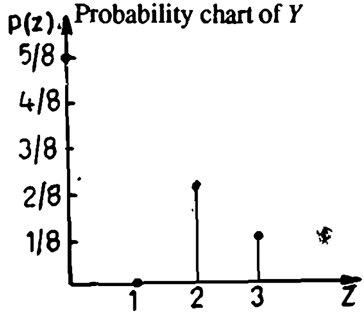
This is obvious from table 1.

(iii) From table 1 , we have

Table 4

Probability Table of

Values of Z, (z)	0	1	2	3
$p(z)$	5/8	0	2/8	1/8



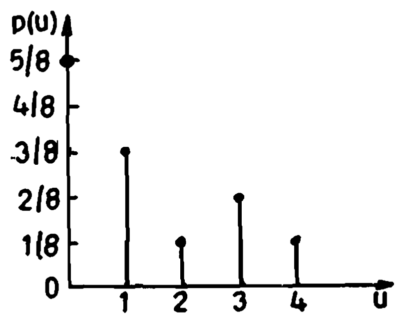
Probability chart of Z

(iv) Let $U = X + Y$. From table 1, we get

Table 5

Probability Table of U

Values of U, (u)	0	1	2	3	4
$p(u)$	1/8	3/8	1/8	2/8	1/8



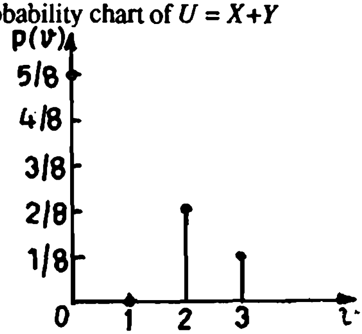
Probability chart of $U = X+Y$

(v) Let $V = XY$

Table 6

Probability Table of V

Values of V, (v)	0	1	2	3
$p(v)$	5/8	0	2/8	1/8



Probability chart of $V = XY$

Example 5·2. *A random variable X has the following probability distribution :*

x :	0	1	2	3	4	5	6	7
$p(x)$:	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2+k$

(i) Find k, (ii) Evaluate $P(X<6)$, $P(X \geq 6)$, *and* $P(0<X<5)$, *(iii) If* $P(X \leq c) > \frac{1}{2}$, *find the minimum value of c, and (iv) Determine the distribution function of X.* **[Madurai Univ. B.Sc., Oct. 1988]**

Solution. Since $\sum_{x=0}^{7} p(x) = 1$, we have

$$\Rightarrow \quad k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$\Rightarrow \quad 10k^2 + 9k - 1 = 0$$

$$\Rightarrow \quad (10k-1)(k+1) = 0 \Rightarrow k = 1/10$$

[$\because k = -1$, is rejected, since probability canot be negative.]

(*ii*) $P(X<6) = P(X=0) + P(X=1) + \ldots + P(X=5)$

$$= \frac{1}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10} + \frac{1}{100} = \frac{81}{100}$$

$$P(X \geq 6) = 1 - P(X<6) = \frac{19}{100}$$

$P(0<X<5) = P(X=1) + P(X=2) + P(X=3) + P(X=4) = 8k = 4/5$

(*iii*) $P(X \leq c) > \frac{1}{2}$. By trial, we get $c = 4$.

(*iv*)

X	$F_X(x) = P(X \leq x)$
0	0
1	$k = 1/10$
2	$3k = 3/10$
3	$5k = 5/10$
4	$8k = 4/5$
5	$8k + k^2 = 81/100$
6	$8k + 3k^2 = 83/100$
7	$9k + 10k^2 = 1$

EXERCISE 5 (b)

1. (*a*) A student is to match three historical events (Mahatma Gandhi's Birthday, India's freedom, and First World War) with three years (1947, 1914, 1896). If he guesses with no knowledge of the correct answers, what is the probability distribution of the number of answers he gets correctly ?

(*b*) From a lot of 10 items containing 3 defectives, a sample of 4 items is drawn at random. Let the random variable X denote the number of defective items in the sample. Answer the following when the sample is drawn without replacement.

(*i*) Find the probability distribution of X,

(*ii*) Find $P(X \le 1)$, $P(X < 1)$ and $P(0 < X < 2)$

Ans. (*a*)

x	0	1	2	3
$p(x)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{6}$

(*b*) (*i*)

x	0	1	2	3
$p(x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

(*ii*) 2/3, 5/6, 1/2

2. (*a*) A random variable X can take all non-negative integral values, and the probability that X takes the value r is proportional to α^r $(0 < \alpha < 1)$. Find $P(X = 0)$. **[Calcutta Univ. B.Sc. 1987]**

Ans. $P(X = r) = A\alpha^r$; $r = 0, 1, 2, \ldots$; $A = 1 - \alpha$; $P(X = 0) = A = 1 - \alpha$

(*b*) Suppose that the random variable X has possible values 1, 2, 3, ... and $P(X = j) = 1/2^j$, $j = 1, 2, \ldots$ (*i*) Compute $P(X$ *is even*$)$, (*ii*) Compute $P(X \ge 5)$, and (*iii*) Compute $P(X$ *is divisible by* $3)$.

Ans. (*i*) 1/3, (*ii*) 1/16, and (*iii*) 1/7

3. (*a*) Let X be a random variable such that

$$P(X = -2) = P(X = -1),\ P(X = 2) = P(X = 1) \text{ and}$$
$$P(X > 0) = P(X < 0) = P(X = 0).$$

Obtain the probability mass function of X and its distribution function.

Ans.

X	:	-2	-1	0	1	2
$p(x)$	:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$
$F(x)$	:	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

(*b*) A random variable X assumes the values $-3, -2, -1, 0, 1, 2, 3$ such that

$$P(X = -3) = P(X = -2) = P(X = -1),$$
$$P(X = 1) = P(X = 2) = P(X = 3),$$
and $$P(X = 0) = P(X > 0) = P(X < 0),$$

Obtain the probability mass function of X and its distribution function, and find further the probability mass function of $Y = 2X^2 + 3X + 4$.

[Poona Univ. B.Sc., March 1991]

Ans.

X	:	-3	-2	-1	0	1	2	3
$p(x)$	:	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
Y	:	13	6	3	4	9	18	31
$p(y)$	:	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

4. *(a)* A random variable X has the following probability function :

Values of X, x :	-2	-1	0	1	2	3
$p(x)$:	$0{\cdot}1$	k	$0{\cdot}2$	$2k$	$0{\cdot}3$	k

(i) Find the value of k, and calculate mean and variance.

(ii) Construct the c.d.f. $F(X)$ and draw its graph.

Ans. *(i)* 0·1, 0·8 and 2·16, *(ii)* $F(X) = 0{\cdot}1, 0{\cdot}2, 0{\cdot}4, 0{\cdot}6, 0{\cdot}9, 1{\cdot}0$

(b) Given the probability function

x	0	1	2	3
$p(x)$	0·1	0·3	0·5	0·1

Let $Y = X^2 + 2X$, then find *(i)* the probability function of Y, *(ii)* mean and variance of Y.

Ans. *(i)*

y	0	3	8	15
$p(y)$	0·1	0·3	0·5	0·1

, *(ii)* 6·4 ,16·24

5. A random variable X has the following probability distribution :

Values of X, x	0	1	2	3	4	5	6	7	8
$p(x)$	a	$3a$	$5a$	$7a$	$9a$	$11a$	$13a$	$15a$	$17a$

(i) Determine the value of a.

(ii) Find $P(X < 3)$, $P(X \geq 3)$, $P(0 < X < 5)$.

(iii) What is the smallest value of x for which $P(X \leq x) > 0{\cdot}5$? and

(iv) Find out the distribution function of X ?

Ans. *(i)* $a = 1/81$, *(ii)* 9/81, 72/81, 24/81, *(iii)* 6

(iv) x	0	1	2	3	4	5	6	7	8
$F(x)$	a	$4a$	$9a$	$16a$	$25a$	$36a$	$49a$	$64a$	$81a$

6. *(a)* Let $p(x)$ be the probability function of a discrete random variable X which assumes the values x_1, x_2, x_3, x_4, such that $2p(x_1) = 3p(x_2) = p(x_3) = 5p(x_4)$. Find probability distribution and cumulative probability distribution of X. **(Sardar Patel Univ. B.Sc. 1987)**

Ans.

x	x_1	x_2	x_3	x_4
$p(x)$	15/16	10/16	30/16	6/16

(b) The following is the distribution function of a discrete random variable X :

x	-3	-1	0	1	2	3	5	8
$f(x)$	0·10	0·30	0·45	0·5	0·75	0·90	0·95	1·00

(i) Find the probability distribution of X.

(ii) Find $P(X \text{ is even})$ and $P(1 \leq X \leq 8)$.

(iii) Find $P(X = -3 \mid X < 0)$ and $P(X \geq 3 \mid X > 0)$.

[**Ans.** *(ii)* 0·30, 0·55, *(iii)* 1/3, 5/11].

7. If $$p(x) = \frac{x}{15};\ x = 1, 2, 3, 4, 5$$
$$= 0, \text{ elsewhere}$$
Find *(i)* $P\{X = 1 \text{ or } 2\}$, and *(ii)* $P\left\{\frac{1}{2} < X < \frac{5}{2} \mid X > 1\right\}$

[Allahabad Univ. B.Sc., April 1992]

Hint. *(i)* $P\{X = 1 \text{ or } 2\} = P(X = 1) + P(X = 2) = \frac{1}{15} + \frac{2}{15} = \frac{1}{5}$

(ii) $$P\left\{\frac{1}{2} < X < \frac{5}{2} \mid X > 1\right\} = \frac{P\left\{\left(\frac{1}{2} < X < \frac{5}{2}\right) \cap X > 1\right\}}{P(X > 1)}$$
$$= \frac{P\{(X = 1 \text{ or } 2) \cap X > 1\}}{P(X > 1)} = \frac{P(X = 2)}{1 - P(X = 1)} = \frac{2/15}{1 - (1/15)} = \frac{1}{7}$$

8. The probability mass function of a random variable X is zero except at the points $x = 0, 1, 2$. At these points it has the values $p(0) = 3c^3$, $p(1) = 4c - 10c^2$ and $p(2) = 5c - 1$ for some $c > 0$.

(i) Determine the value of c.

(ii) Compute the following probabilities, $P(X < 2)$ and $P(1 < X \leq 2)$.

(iii) Describe the distribution function and draw its graph.

(iv) Find the largest x such that $F(x) < 1/2$.

(v) Find the smallest x such that $F(x) \geq 1/3$. **[Poona Univ. B.Sc., 1987]**

Ans. *(i)* $\frac{1}{3}$, *(ii)* $\frac{1}{3}, \frac{2}{3}$, *(iv)* 1, *(v)* 1.

9. *(a)* Suppose that the random variable X assumes three values 0,1 and 2 with probabilities $\frac{1}{3}$, $\frac{1}{6}$ and $\frac{1}{2}$ respectively. Obtain the distribution function of X..

[Gujarat Univ. B.Sc., 1992]

(b) Given that $f(x) = k(1/2)^x$ is a probability distribution for a random variable which can take on the values $x = 0, 1, 2, 3, 4, 5, 6$, find k and find an expression for the corresponding cumulative probabilities $F(x)$.

[Nagpur Univ. B.Sc., 1987)

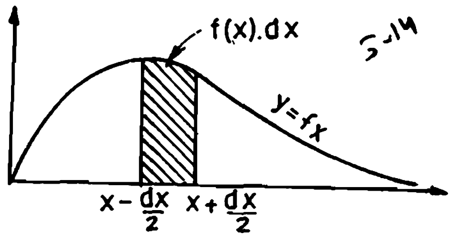
5·4. Continuous Random Variable. A random variable X is said to be continuous if it can take all possible values between certain limits. *In other words, a random variable is said to be continuous when its different values cannot be put in 1-1 correspondence with a set of positive integers.*

A continuous random variable is a random variable that (at least conceptually) can be measured to any desired degree of accuracy. Examples of continuous random variables are age, height, weight etc.

5·4·1. Probability Density Function *(Concept and Definition).* Consider the small interval $(x, x + dx)$ of length dx round the point x. Let $f(x)$ be any continuous function of x so that $f(x)\,dx$ represents the probability that X falls in the infinitesimal interval $(x, x+dx)$. Symbolically

$$P(x \le X \le x + dx) = f_X(x)\,dx \qquad \dots (5\cdot5)$$

In the figure, $f(x)\,dx$ represents the area bounded by the curve $y = f(x)$, x–axis and the ordinates at the points x and $x + dx$. The function $f_X(x)$ so defined is known as *probability density function or simply density function of random variable X and is usually abbreviated as p.d.f.* The expression, $f(x)\,dx$, usually written as $dF(x)$, is known as the *probability differential* and the curve $y = f(x)$ is known as the *probability density curve* or simply *probability curve.*



Definition. p.d.f. $f_X(x)$ of the *r.v.* X is defined as :

$$f_X(x) = \lim_{\delta x \to 0} \frac{P(x \le X \le x + \delta x)}{\delta x} \qquad \dots (5\cdot5\,a)$$

The probability for a variate value to lie in the interval dx is $f(x)\,dx$ and hence the probability for a variate value to fall in the finite interval $[\alpha, \beta]$ is :

$$P(\alpha \le X \le \beta) = \int_{\alpha}^{\beta} f(x)\,dx \qquad \dots (5\cdot5\,b)$$

which represents the area between the curve $y = f(x)$, x–axis and the ordinates at $x = \alpha$ and $x = \beta$. Further since total probability is unity, we have $\int_a^b f(x)\,dx = 1$, where $[a, b]$ is the range of the random variable X. The range of the variable may be finite or infinite.

The probability density function (*p.d.f.*) of a random variable (*r.v.*) X usually denoted by $f_X(x)$ or simply by $f(x)$ has the following obvious properties

(i) $f(x) \ge 0, \; -\infty < x < \infty$ $\qquad \dots (5\cdot5\,c)$

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ $\qquad \dots (5\cdot5\,d)$

(iii) The probability $P(E)$ given by

$$P(E) = \int_E f(x)\,dx \qquad \dots (5\cdot5\,e)$$

is well defined for any event E.

Important Remark. In case of discrete random variable, the probability at a point, *i.e.*, $P(x = c)$ is not zero for some fixed c. However, in case of continuous random variables the probability at a point is always zero, *i.e.*, $P(x = c) = 0$ for all possible values of c. This follows directly from (5·5 *b*) by taking $\alpha = \beta = c$.

This also agrees with our discussion earlier that $P(E) = 0$ does not imply that the event E is null or impossible event. This property of continuous r.v., *viz.*,

$$P(X = c) = 0, \ \forall \ c \qquad \ldots (5{\cdot}5 f)$$

leads us to the following important result :

$$P(\alpha \le X \le \beta) = P(\alpha \le X < \beta) = P(\alpha < X \le \beta) = P(\alpha < X < \beta) \quad \ldots (5{\cdot}5 g)$$

i.e., in case of continuous r.v., it does matter whether we include the end points of the interval from α to β.

However, this result is in general not true for discrete random variables.

5·4·2. Various Measures of Central Tendency, Dispersion, Skewness, and Kurtosis for Continuous Probability Distribution. The formulae for these measures in case of discrete frequency distribution can be easily extended to the case of continuous probability distribution by simply replacing $p_i = f_i/N$ by $f(x)\,dx$, x_i by x and the summation over 'i' by integration over the specified range of the variable X.

Let $f_X(x)$ *or* $f(x)$ be the *p.d.f.* of a random variable X where X is defined from a to b. Then

(i) *Arithmetic mean* $= \int_a^b x\, f(x)\, dx$ ...(5·6)

(ii) Harmonic mean. Harmonic mean H is given by

$$\frac{1}{H} = \int_a^b \left(\frac{1}{x}\right) f(x)\, dx \qquad \ldots(5{\cdot}6\ a)$$

(iii) Geometric mean. Geometric mean G is given by

$$\log G = \int_a^b \log x f(x)\, dx \qquad \ldots(5{\cdot}6\ b)$$

(iv) μ_r' (about origin) $= \int_a^b x^r f(x)\, dx$ ...(5·7)

μ_r' (about the point $x = A$) $= \int_a^b (x - A)^r f(x)\, dx$ ...(5·7 a)

and μ_r (about mean) $= \int_a^b (x - mean)^r f(x)\, dx$ ...(5·7 b)

In particular, from (5·7), we have

μ_1' (about origin) $=$ *Mean* $= \int_a^b x\, f(x)\, dx$

and $\mu_2' = \int_a^b x^2 f(x)\, dx$

Hence $\mu_2 = \mu_2' - \mu_1'^2 = \int_a^b x^2 f(x)\, dx - \left(\int_a^b x f(x)\, dx\right)^2$ ... (5·7 c)

From (5·7), on putting $r=3$ and 4 respectively, we get the values of μ_3' and μ_4' and consequently the moments about mean can be obtained by using the relations :

and
$$\left.\begin{aligned}\mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \\ \mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4\end{aligned}\right\} \quad \ldots (5\cdot 7\ d)$$

and hence β_1 and β_2 can be computed.

(v) Median. Median is the point which divides the entire distribution in two equal parts. In case of continuous distribution, median is the point which divides the total area into two equal parts. Thus if M is the median, then

$$\int_a^M f(x)\,dx = \int_M^b f(x)\,dx = \frac{1}{2} \quad \ldots (5\cdot 8)$$

Thus solving

$$\int_a^M f(x)\,dx = \frac{1}{2} \quad \textit{or} \quad \int_M^b f(x)\,dx = \frac{1}{2} \quad \ldots (5\cdot 8\ a)$$

for M, we get the value of median.

(vi) Mean Deviation. Mean deviation about the mean μ_1' is given by

$$M.D. = \int_a^b |\,x - mean\,|\,f(x)\,dx \quad \ldots (5\cdot 9)$$

(vii) Quartiles and Deciles. Q_1 and Q_3 are given by the equations

$$\int_a^{Q_1} f(x)\,dx = \frac{1}{4} \quad \text{and} \quad \int_a^{Q_3} f(x)\,dx = \frac{3}{4} \quad \ldots (5\cdot 10)$$

D_i, i th decile is given by

$$\int_a^{D_i} f(x)\,dx = \frac{i}{10} \quad \ldots (5\cdot 10\ a)$$

(viii) Mode. Mode is the value of x for which $f(x)$ is maximum. Mode is thus the solution of

$$f'(x) = 0 \text{ and } f''(x) < 0 \quad \ldots (5\cdot 11)$$

provided it lies in $[a,b]$.

Example 5·3. *The diameter of an electric cable, say X, is assumed to be a continuous random variable with p.d.f.* $f(x) = 6x(1-x)$, $0 \le x \le 1$.

(i) Check that above is p.d.f.,

(ii) Determine a number b such that $P(X < b) = P(X > b)$

[Aligarh Univ. B.Sc.(Hons).1990]

Solution. Obviously, for $0 \le x \le 1$, $f(x) \ge 0$

Now
$$\int_0^1 f(x)\,dx = 6\int_0^1 x(1-x)\,dx$$
$$= 6\int_0^1 (x - x^2)\,dx = 6\left|\frac{x^2}{2} - \frac{x^3}{3}\right|_0^1 = 1$$

Hence $f(x)$ is the *p.d.f.* of *r.v.* X

(ii) $$P(X < b) = P(X > b) \quad \ldots(*)$$

$$\Rightarrow \qquad \int_0^b f(x)\,dx = \int_b^1 f(x)\,dx$$

$$\Rightarrow \qquad 6\int_0^b x(1-x)\,dx = 6\int_b^1 x(1-x)\,dx$$

$$\Rightarrow \qquad \left| \frac{x^2}{2} - \frac{x^3}{3} \right|_0^b = \left| \frac{x^2}{2} - \frac{x^3}{3} \right|_b^1$$

$$\Rightarrow \qquad \left(\frac{b^2}{2} - \frac{b^3}{3} \right) = \left[\left(\frac{1}{2} - \frac{1}{3} \right) - \left(\frac{b^2}{2} - \frac{b^3}{3} \right) \right]$$

$$\Rightarrow \qquad 3b^2 - 2b^3 = [\,1 - 3b^2 + 2b^3\,]$$

$$\Rightarrow \qquad 4b^3 - 6b^2 + 1 = 0$$

$$(2b-1)(2b^2 - 2b - 1) = 0$$

$$\Rightarrow \qquad 2b - 1 = 0 \quad or \quad 2b^2 - 2b - 1 = 0$$

Hence $b = 1/2$ is the only real value lying between 0 and 1 and satisfying (*).

Example 5·4. *A continuous random variable* X *has a p.d.f.*

$$f(x) = 3x^2,\ 0 \le x \le 1. \quad \textit{Find a and b such that}$$

(i) $P\{X \le a\} = P\{X > a\}$, and

(ii) $P\{X > b\} = 0{\cdot}05$. [Calicut Univ. B.Sc., Sept. 1988]

Solution. *(i)* Since $P(X \le a) = P(X > a)$,
each must be equal to $1/2$, because total probability is always one.

$$\therefore \qquad P(X \le a) = \frac{1}{2} \quad \Rightarrow \quad \int_0^a f(x)\,dx = \frac{1}{2}$$

$$\Rightarrow \qquad 3\int_0^a x^2\,dx = \frac{1}{2} \quad \Rightarrow \quad 3\left| \frac{x^3}{3} \right|_0^a = \frac{1}{2}$$

$$\Rightarrow \qquad a^3 = \frac{1}{2} \quad \Rightarrow \quad a = \left(\frac{1}{2} \right)^{\frac{1}{3}}$$

(ii)
$$P(X > b) = 0{\cdot}05 \quad \Rightarrow \quad \int_b^1 f(x)\,dx = 0{\cdot}05$$

$$\Rightarrow \qquad 3\left| \frac{x^3}{3} \right|_b^1 = \frac{1}{20} \quad \Rightarrow \quad 1 - b^3 = \frac{1}{20}$$

$$\Rightarrow \qquad b^3 = \frac{19}{20} \quad \Rightarrow \quad b = \left(\frac{19}{20} \right)^{\frac{1}{3}}.$$

Example 5·5. *Let* X *be a continuous random variate with p.d.f.*

$$\begin{aligned} f(x) &= ax, \quad 0 \le x \le 1 \\ &= a, \quad 1 \le x \le 2 \\ &= -ax + 3a, \quad 2 \le x \le 3 \\ &= 0, \textit{ elsewhere} \end{aligned}$$

(i) Determine the constant a.

(ii) Compute $P(X \leq 1\cdot5)$. **[Sardar Patel Univ. B.Sc., Nov.1988]**

Solution. *(i)* Constant 'a' is determined from the consideration that total probability is unity, *i.e.,*

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\Rightarrow \int_{-\infty}^{0} f(x)\,dx + \int_0^1 f(x)\,dx + \int_1^2 f(x)\,dx + \int_2^3 f(x)\,dx + \int_3^{\infty} f(x)\,dx = 1$$

$$\Rightarrow \int_0^1 ax\,dx + \int_1^2 a.dx + \int_2^3 (-ax+3a)\,dx = 1$$

$$\Rightarrow a\left|\frac{x^2}{2}\right|_0^1 + a\left|x\right|_1^2 + a\left|-\frac{x^2}{2}+3x\right|_2^3 = 1$$

$$\Rightarrow \frac{a}{2} + a + a\left[\left(-\frac{9}{2}+9\right) - (-2+6)\right] = 1$$

$$\Rightarrow \frac{a}{2} + a + \frac{a}{2} = 1 \quad \Rightarrow \quad 2a = 1 \quad \Rightarrow \quad a = \frac{1}{2}$$

(ii) $$P(X \leq 1\cdot5) = \int_{-\infty}^{1\cdot5} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_0^1 f(x)\,dx + \int_1^{1\cdot5} f(x)\,dx$$

$$= a\int_0^1 x\,dx + \int_1^{1\cdot5} a.dx$$

$$= a\left|\frac{x^2}{2}\right|_0^1 + a\left|x\right|_1^{1\cdot5} = \frac{a}{2} + 0\cdot5\,a$$

$$= a = \frac{1}{2} \qquad \left[\because a = \frac{1}{2},\ \textit{Part (i)}\right]$$

Example 5·6. *A probability curve $y = f(x)$ has a range from 0 to ∞. If $f(x) = e^{-x}$, find the mean and variance and the third moment about mean.*

[Andhra Univ. B.Sc. 1988; Delhi Univ. B.Sc. Sept. 1987]

Solution.

$$\mu_r' \text{ (rth moment about origin)} = \int_0^{\infty} x^r f(x)\,dx$$

$$= \int_0^{\infty} x^r e^{-x}\,dx = \Gamma(r+1) = r\,!$$

(Using Gamma Integral)

Substituting $r = 1, 2$ and 3 successively, we get

$$\text{Mean} = \mu_1' = 1\,! = 1,\ \mu_2' = 2\,! = 2,\ \mu_3' = 3\,! = 6$$

Hence variance $= \mu_2 = \mu_2' - \mu_1'^2 = 2 - 1 = 1$

and $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 6 - 3\times2 + 2 = 2$

Example 5·7. *In a continuous distribution whose relative frequency density is given by*

$$f(x) = y_0 \,.\, x(2-x),\ 0 \le x \le 2,$$

find mean, variance, β_1, and β_2 and hence show that the distribution is symmetrical. Also (i) find mean deviation about mean and (ii) show that for this distribution $\mu_{2n+1} = 0$, (iii) find the mode, harmonic mean and median.

[Delhi Univ. B.Sc.(Stat. Hons.), 1992; B.Sc., Oct. 1992]

Solution. Since total probability is unity, we have

$$\int_0^2 f(x)\,dx = 1$$

$$\Rightarrow \qquad y_0 \int_0^2 x(2-x)\,dx = 1 \quad \Rightarrow \quad y_0 = 3/4$$

$$\therefore \qquad f(x) = \frac{3}{4} x(2-x)$$

$$\mu_r' = \int_0^2 x^r f(x)\,dx = \frac{3}{4}\int_0^2 x^{r+1}(2-x)\,dx = \frac{3 \cdot 2^{r+1}}{(r+2)(r+3)}$$

In particular

$$\text{Mean} = \mu_1' = \frac{3 \cdot 2^2}{3 \cdot 4} = 1, \quad \mu_2' = \frac{3 \cdot 2^3}{4 \cdot 5} = \frac{6}{5},$$

$$\mu_3' = \frac{3 \cdot 2^4}{5 \cdot 6} = \frac{8}{5}, \quad \text{and} \quad \mu_4' = \frac{3 \cdot 2^5}{6 \cdot 7} = \frac{16}{7}$$

Hence varience $= \mu_2 = \mu_2' - \mu_1'^2 = \frac{6}{5} - 1 = \frac{1}{5}$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \frac{8}{5} - 3 \,.\, \frac{6}{5} \,.\, 1 + 2 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = \frac{16}{7} - 4 \,.\, \frac{8}{5} \,.\, 1 + 6 \,.\, \frac{6}{5} \,.\, 1 - 3 \,.\, 1 = \frac{3}{35}$$

$$\therefore \qquad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3/35}{(1/5)^2} = \frac{15}{7}$$

Since $\beta_1 = 0$, the distributior is *symmetrical.*

Mean deviation about mean

$$= \int_0^2 |x-1|\, f(x)\,dx$$

$$= \int_0^1 |x-1|\, f(x)\,dx + \int_1^2 |x-1|\, f(x)\,dx$$

$$= \frac{3}{4}\left[\int_0^1 (1-x)\,x(2-x)\,dx + \int_1^2 (x-1)\,x(2-x)\,dx\right]$$

$$= \frac{3}{4}\left[\int_0^1 (2x - 3x^2 + x^3)\,dx + \int_1^2 (3x^2 - x^3 - 2x)\,dx\right]$$

$$= \frac{3}{4}\left[\left| x^2 - \frac{3.x^3}{3} + \frac{x^4}{4}\right|_0^1 + \left|3.\frac{x^3}{3} - \frac{x^4}{4} - \frac{2x^2}{2}\right|_1^2\right] = \frac{3}{8}$$

$$\mu_{2n+1} = \int_0^2 (x - mean)^{2n+1} f(x)\,dx$$

$$= \frac{3}{4}\int_0^2 (x-1)^{2n+1} x(2-x)\,dx$$

$$= \frac{3}{4}\int_{-1}^1 t^{2n+1}(t+1)(1-t)\,dt \qquad (x - 1 = t)$$

$$= \frac{3}{4}\int_{-1}^1 t^{2n+1}(1-t^2)\,dt$$

Since t^{2n+1} 1 is an odd function of t and $(1 - t^2)$ is an even function of t, the integrand $t^{2n+1}(1-t^2)$ is an odd function of t.

Hence $\mu_{2n+1} = 0.$

Now $f'(x) = \frac{3}{4}(2 - 2x) = 0 \Rightarrow x = 1$

and $f''(x) = \frac{3}{4}(-2) = -\frac{3}{2} < 0$

Hence mode = 1

Harmonic mean H is given by

$$\frac{1}{H} = \int_0^2 \frac{1}{x} f(x)\,dx$$

$$= \frac{3}{4}\int_0^2 (2-x)\,dx = \frac{3}{2}$$

$$\Rightarrow \qquad H = \frac{2}{3}$$

If M is the median, then

$$\int_0^M f(x)\,dx = \frac{1}{2}$$

$$\Rightarrow \qquad \frac{3}{4}\int_0^M x(2-x)\,dx = \frac{1}{2}$$

$$\Rightarrow \qquad \left| x^2 - \frac{x^3}{3}\right|_0^M = \frac{2}{3}$$

$$\Rightarrow \qquad 3M^2 - M^3 = 2$$

$$\Rightarrow \qquad M^3 - 3M^2 + 2 = 0$$

$$\Rightarrow \qquad (M-1)(M^2 - 2M - 2) = 0$$

The only value of M lying in $[0, 2]$ is $M = 1$. Hence median is 1.

Aliter. Since we have proved that distribution is symmetrical,

$$\text{Mode} = \text{Median} = \text{Mean} = 1$$

Example 5·8. *The elementary probability law of a continuous random variable X is*

$$f(x) = y_o e^{-b(x-a)}, \quad a \le x < \infty, \quad b > 0$$

where a, b and y_o are constants.

Show that $y_o = b = 1/\sigma$ and $a = m - \sigma$, where m and σ are respectively the mean and standard deviation of the distribution. Show also that $\beta_1 = 4$ and $\beta_2 = 9$. **[Gauhati Univ. B.Sc., 1992]**

Solution. Since total probability is unity,

$$\int_a^\infty f(x)\,dx = 1 \quad \Rightarrow \quad y_o \int_a^\infty e^{-b(x-a)}\,dx = 1$$

$$\Rightarrow \quad y_o \left| \frac{e^{-b(x-a)}}{-b} \right|_a^\infty = 1 \quad \Rightarrow \quad y_o \frac{1}{b} = 1, \ (b > 0)$$

$$\Rightarrow \quad y_o = b$$

μ_r' (rth moment about the point '$x = a$')

$$= \int_a^\infty (x-a)^r f(x)\,dx = b \int_a^\infty (x-a)^r e^{-b(x-a)}\,dx$$

$$= b \int_0^\infty t^r e^{-bt}\,dt \qquad \text{[On putting } x - a = t \text{]}$$

$$= b\, \frac{\Gamma(r+1)}{b^{r+1}} = \frac{r\,!}{b^r} \qquad \text{[Using Gamma Integral]}$$

In particular

$$\mu_1' = 1/b, \ \mu_2' = 2/b^2, \ \mu_3' = 6/b^3, \ \mu_4' = 24/b^4$$

$$\therefore \quad m = Mean = a + \mu_1' = a + (1/b)$$

$$\text{and} \quad \sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = 1/b^2$$

$$\Rightarrow \quad \sigma = \frac{1}{b} \ \text{ and } \ m = a + \frac{1}{b} = a + \sigma$$

$$\text{Hence} \quad y_o = b = \frac{1}{\sigma} \ \text{ and } \ a = m - \sigma$$

$$\text{Also} \quad \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \frac{1}{b^3}(6 - 3\cdot2 + 2) = \frac{2}{b^3} = 2\sigma^3$$

$$\text{and} \quad \mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= \frac{1}{b^4}(24 - 4.6.1 + 6.2.1 - 3) = \frac{9}{b^4} = 9\sigma^4$$

Hence $\beta_1 = \mu_3^2/\mu_2^3 = 4\sigma^6/\sigma^6 = 4$ and $\beta_2 = \mu_4/\mu_2^2 = 9\sigma^4/\sigma^4 = 9$

Example 5·9. *For the following probability distribution*

$$dF = y_o \,.\, e^{-|x|}\, dx\,,\quad -\infty < x < \infty$$

show that $y_o = \frac{1}{2}$, $\mu_1' = 0$, $\sigma = \sqrt{2}$ *and mean deviation about mean = 1.*

Solution. We have $\int_{-\infty}^{\infty} f(x)\, dx = 1$

$$\Rightarrow\quad y_o \int_{-\infty}^{\infty} e^{-|x|}\, dx = 1 \quad\Rightarrow\quad 2y_o \int_0^{\infty} e^{-|x|}\, dx = 1,$$

(since $e^{-|x|}$ is an even function of x)

$$\Rightarrow\quad 2y_o \int_0^{\infty} e^{-x}\, dx = 1,$$

(since in $0 \le x < \infty$, $|x| = x$)

$$\Rightarrow\quad 2y_o \left| \frac{e^{-x}}{-1} \right|_0^{\infty} = 1 \quad\Rightarrow\quad 2y_o = 1,\ i.e.,\ y_o = \frac{1}{2}$$

$$\mu_1' \text{ (about origin)} = \int_{-\infty}^{\infty} x\, f(x)\, dx = \frac{1}{2}\int_{-\infty}^{\infty} x\, e^{-|x|}\, dx$$

$$= 0,$$

(since the integrand $x \,.\, e^{-|x|}$ is an odd function of x)

$$\mu_2' = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-|x|}\, dx$$

$$= \frac{1}{2}\, 2 \int_0^{\infty} x^2 e^{-|x|}\, dx$$

[since the integrand $x^2 e^{-|x|}$ is an even function of x]

$$\therefore\quad \mu_2' = \int_0^{\infty} x^2 e^{-x}\, dx = \Gamma(3) \quad \text{(on using Gamma Integral)}$$

$$\Rightarrow\quad \mu_2' = 2\,! = 2$$

Now $\sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = 2$

$$\text{M.D. about mean} = \int_{-\infty}^{\infty} |x - mean|\, f(x)\, dx$$

$$= \frac{1}{2}\int_{-\infty}^{\infty} |x|\, e^{-|x|}\, dx \qquad (\because \text{Mean} = \mu_1' = 0)$$

$$= \frac{1}{2} \cdot 2 \int_0^{\infty} |x|\, e^{-|x|}\, dx$$

$$= \int_0^{\infty} x e^{-x}\, dx = \Gamma(2) = 1$$

Example 5·10. *A random variable X has the probability law :*

$$dF(x) = \frac{x}{b^2} \cdot e^{-x^2/2b^2}\, dx, \quad 0 \le x < \infty$$

Find the distance between the quartiles and show that the ratio of this distance to the standard devation of X is independent of the patameter 'b'.

Solution. If Q_1 and Q_3 are the first and third quartiles respectively, we have

$$\int_0^{Q_1} f(x)\, dx = \frac{1}{4} \quad \Rightarrow \quad \frac{1}{b^2} \int_0^{Q_1} x\, e^{-x^2/2b^2}\, dx = \frac{1}{4}$$

Put $\quad y = \dfrac{x^2}{2b^2} \quad$ *then* $\quad dy = \dfrac{x}{b^2}\, dx$

$$\therefore \quad \int_0^{Q_1^2/2b^2} e^{-y}\, dy = \frac{1}{4} \quad \Rightarrow \quad \left| \frac{e^{-y}}{-1} \right|_0^{Q_1^2/2b^2} = \frac{1}{4}$$

$$\Rightarrow \quad 1 - e^{-Q_1^2/2b^2} = \frac{1}{4} \quad \Rightarrow \quad e^{-Q_1^2/2b^2} = \frac{3}{4}$$

$$\Rightarrow \quad Q_1 = \sqrt{2}\, b \sqrt{\log (4/3)}$$

Again we have $\int_0^{Q_3} f(x)\, dx = \dfrac{3}{4}$ which, on proceeding similarly, will give

$$1 - e^{-Q_3^2/2b^2} = 3/4 \quad \Rightarrow \quad e^{-Q_3^2/2b^2} = 1/4$$

$$\Rightarrow \quad Q_3 = \sqrt{2}\, b \sqrt{\log (4)}$$

The distance between the quartiles is given by

$$Q_3 - Q_1 = \sqrt{2}\, b \left[\sqrt{\log 4} - \sqrt{\log (4/3)}\right]$$

$$\mu_1' = \int_0^{\infty} x\, f(x)\, dx = \int_0^{\infty} x\, \frac{x}{b^2}\, e^{-x^2/2b^2}\, dx$$

$$= \int_0^{\infty} \sqrt{2}\, b y^{1/2}\, e^{-y}\, dy \qquad \left(y = \frac{x^2}{2b^2}\right)$$

$$= \sqrt{2}\, b \int_0^{\infty} e^{-y}\, y^{(3/2)-1}\, dy$$

$$= \sqrt{2}\, b\, \Gamma\left(\frac{3}{2}\right) = \sqrt{2} . b . \frac{1}{2}\, \Gamma\left(\frac{1}{2}\right) = \sqrt{2}\, b \frac{\sqrt{\pi}}{2} = b\sqrt{(\pi/2)}$$

$$\mu_2' = \int_0^{\infty} x^2 f(x)\, dx = \int_0^{\infty} x^2\, \frac{x}{b^2}\, e^{-x^2/2b^2}\, dx$$

$$= 2b^2 \int_0^{\infty} y\, e^{-y}\, dy \qquad \left(y = \frac{x^2}{2b^2}\right)$$

$$= 2b^2\, \Gamma(2) = 2b^2 . 1\,! = 2b^2$$

$$\therefore \quad \sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = 2b^2 - b^2 \cdot \frac{\pi}{2} = b^2\left(2 - \frac{\pi}{2}\right)$$

$$\Rightarrow \quad \sigma = b\sqrt{2 - (\pi/2)}$$

Hence $$\frac{Q_3 - Q_1}{\sigma} = \frac{\sqrt{2}\,[\sqrt{\log 4} - \sqrt{\log(4/3)}\,]}{\sqrt{2 - (\pi/2)}},$$

which is independent of the parameter 'b'.

Example 5·11. *Prove that the geometric mean G of the distribution*

$$dF = 6(2 - x)(x - 1)\,dx, \quad 1 \le x \le 2$$

is given by $6 \log(16G) = 19$. **[Kanpur Univ. B.Sc., Oct. 1992]**

Solution. By definition, we have

$$\log G = \int_1^2 \log x \; f(x)\,dx = 6\int_1^2 \log x \;(2 - x)(x - 1)\,dx$$

$$= -6\int_1^2 (x^2 - 3x + 2) \log x \,dx$$

Integrating by parts, we get

$$\log G = -6\left[\left|\left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\log x\right|_1^2 - \int_1^2 \left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\frac{1}{x}\,dx\right]$$

$$= -4\log 2 + 6 \times \frac{19}{36} \qquad \text{(on simplification)}$$

$$\therefore \quad \log G + 4\log 2 = \frac{19}{6} \quad \Rightarrow \quad \log G + \log 2^4 = \frac{19}{6}$$

$$\Rightarrow \quad \log G + \log 16 = \frac{19}{6} \quad \Rightarrow \quad \log(16\,G) = \frac{19}{6}$$

$$\Rightarrow \quad 6\log(16 G) = 19$$

Example 5·12. *The time one has to wait for a bus at a downtown bus stop is observed to be random phenomenon (X) with the following probability density function :*

$$\begin{aligned} f_X(x) &= 0, && \text{for } x < 0 \\ &= \tfrac{1}{9}(x + 1), && \text{for } 0 \le x < 1 \\ &= \tfrac{4}{9}\left(x - \tfrac{1}{2}\right), && \text{for } 1 \le x < \tfrac{3}{2} \\ &= \tfrac{4}{9}\left(\tfrac{5}{2} - x\right), && \text{for } \tfrac{3}{2} \le x < 2 \\ &= \tfrac{1}{9}(4 - x), && \text{for } 2 \le x < 3 \\ &= \tfrac{1}{9}, && \text{for } 3 \le x < 6 \end{aligned}$$

$= 0, \qquad for \quad 6 \le x,$

Let the events A and B be defined as follows :

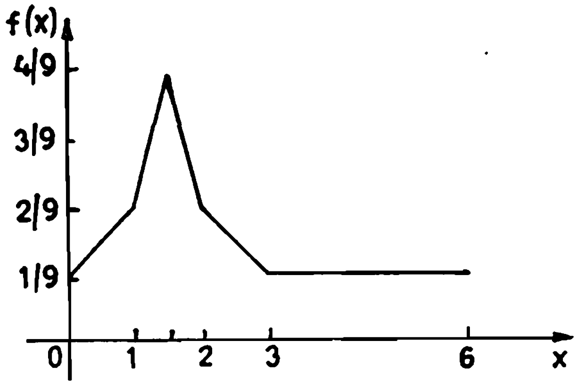
A : One waits between 0 to 2 minutes inclusive:

B : One waits between 0 to 3 minutes inclusive.

(i) Draw the graph of probability density function.

(ii) Show that (a) $P(B \mid A) = \dfrac{2}{3}$, *(b)* $P(\bar{A} \cap \bar{B}) = \dfrac{1}{3}$

Solution. *(i)* The graph of p.d.f. is given below.



(ii) (a) $P(A) = \int_0^2 f(x)\,dx = \int_0^1 \frac{1}{9}(x+1)\,dx + \int_1^{3/2} \frac{4}{9}\left(x - \frac{1}{2}\right)dx + \int_{3/2}^{2} \frac{4}{9}\left(\frac{5}{2} - x\right)dx$

$= \dfrac{1}{2}$ (on simplification)

$P(A \cap B) = P(1 \le X \le 2) = \int_1^2 f(x)\,dx$

$= \int_1^{3/2} \frac{4}{9}\left(x - \frac{1}{2}\right)dx + \int_{3/2}^{2} \frac{4}{9}\left(\frac{5}{2} - x\right)dx$

$= \frac{4}{9}\left[\frac{x^2}{2} - \frac{x}{2}\right]_1^{3/2} + \frac{4}{9}\left[\frac{5}{2}x - \frac{x^2}{2}\right]_{3/2}^{2} = \frac{1}{3}$

(on simplification)

$\therefore \; P(B \mid A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{1/3}{1/2} = \dfrac{2}{3}$

(b) $\bar{A} \cap \bar{B}$ means that waiting time is more than 3 minutes.

$\therefore \; P(\bar{A} \cap \bar{B}) = P(X > 3) = \int_3^\infty f(x)\,dx = \int_3^6 f(x)\,dx + \int_6^\infty f(x)\,dx$

$= \int_3^6 \frac{1}{9}\,dx = \frac{1}{9} \;\Big|\, x \,\Big|_3^6 = \frac{1}{3}$

Example 5·13. *The amount of bread (in hundreds of pounds) X that a certain bakery is able to sell in a day is found to be a numerical valued random phenomenon, with a probability function specified by the probability density function $f(x)$, given by*

$$\begin{aligned} f(x) &= A\,.\,x, && \text{for } 0 \le x < 5 \\ &= A\,(10-x), && \text{for } 5 \le x < 10 \\ &= 0, && \textit{otherwise} \end{aligned}$$

(a) Find the value of A such that $f(x)$ is a probability density function.

(b) What is the probability that the number of pounds of bread that will be sold tomorrow is

(i) more than 500 pounds,

(ii) less than 500 pounds,

(iii) between 250 and 750 pounds? **[Agra Univ. B.Sc., 1989]**

(c) Denoting by A, B, C the events that the pounds of bread sold are as in b (i), b (ii) and b (iii) respectively, find $P(A \mid B), P(A \mid C)$. Are (i) A and B independent events? (ii) Are A and C independent events?

Solution. *(a)* In order that $f(x)$ should be a probability density function

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

i.e.,
$$\int_0^5 A\,x\,dx + \int_5^{10} A\,(10-x)\,dx = 1$$

$$\Rightarrow \qquad A = \frac{1}{25} \qquad \text{(On simplification)}$$

(b) (i) The probability that the number of pounds of bread that will be sold tomorrow is more than 500 pounds, *i.e.,*

$$P(5 \le X \le 10) = \int_5^{10} \frac{1}{25}\,(10-x)\,dx = \frac{1}{25}\left| 10x - \frac{x^2}{2} \right|_5^{10}$$

$$= \frac{1}{25}\left(\frac{25}{2}\right) = \frac{1}{2} = 0{\cdot}5$$

(ii) The probability that the number of pounds of bread that will be sold tomorrow is less than 500 pounds, *i.e.,*

$$P(0 \le X \le 5) = \int_0^5 \frac{1}{25}\,.\,x\,dx = \frac{1}{25}\left| \frac{x^2}{2} \right|_0^5 = \frac{1}{2} = 0{\cdot}5$$

(iii) The required probability is given by

$$P(2{\cdot}5 \le X \le 7{\cdot}5) = \int_{2\cdot5}^{5} \frac{1}{25}\,x\,dx + \int_5^{7\cdot5} \frac{1}{25}\,(10-x)\,dx = \frac{3}{4}$$

(c) The events *A, B* and *C* are given by

$$A : 5 < X \le 10\,; \quad B : 0 \le X < 5\,; \quad C : 2{\cdot}5 < X < 7{\cdot}5$$

Then from parts *b (i), (ii)* and *(iii)*, we have

$$P(A) = 0.5\,,\quad P(B) = 0.5\,,\quad P(C) = \frac{3}{4}$$

The events $A \cap B$ and $A \cap C$ are given by

$$A \cap B = \phi \quad \text{and} \quad A \cap C : 5 < X < 7.5$$

$$\therefore \qquad P(A \cap B) = P(\phi) = 0$$

and $$P(A \cap C) = \int_5^{7.5} f(x)\,dx = \frac{1}{25}\int_5^{7.5} (10 - x)\,dx$$

$$= \frac{1}{25} \times \frac{75}{8} = \frac{3}{8}$$

$$P(A) \,.\, P(C) = \frac{1}{2} \times \frac{3}{4} = \frac{3}{8} = P(A \cap C)$$

$\Rightarrow$ A and C are independent.

Again $P(A) \,.\, P(B) = \dfrac{1}{4} \neq P(A \cap B)$

$\Rightarrow$ A and B are not independent.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = 0$$

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{3/8}{3/4} = \frac{1}{2}$$

Example 5·14. *The mileage C in thousands of miles which car owners get with a certain kind of tyre is a random variable having probability density function*

$$f(x) = \frac{1}{20}\, e^{-x/20}\,, \quad \text{for } x > 0$$

$$= 0\,, \quad \text{for } x \leq 0$$

Find the probabilities that one of these tyres will last

(i) at most 10,000 miles,

(ii) anywhere from 16,000 to 24,000 miles.

(iii) at least 30,000 miles. **(Bombay Univ. B.Sc. 1989)**

Solution. Let *r.v.* X denote the mileage (in '000 miles) with a certain kind of tyre. Then required probability is given by:

(i) $$P(X \leq 10) = \int_0^{10} f(x)\,dx = \frac{1}{20}\int_0^{10} e^{-x/20}\,dx$$

$$= \frac{1}{20} \left| \frac{e^{-x/20}}{-1/20} \right|_0^{10} = 1 - e^{-1/2}$$

$$= 1 - 0.6065 = 0.3935$$

(ii) $P(16 \le X \le 24) = \dfrac{1}{20}\int_{16}^{24} \exp\left(-\dfrac{x}{20}\right)dx = \left| -e^{-x/20} \right|_{16}^{24}$

$= e^{-16/20} - e^{-24/20} = e^{-4/5} - e^{-6/5}$

$= 0{\cdot}4493 - 0{\cdot}3012 = 0{\cdot}1481$

(iii) $P(X \ge 30) = \int_{30}^{\infty} f(x)\,dx = \dfrac{1}{20}\left| \dfrac{e^{-x/20}}{-1/20} \right|_{30}^{\infty}$

$= e^{-1.5} = 0{\cdot}2231$

EXERCISE 5 (c)

1. (a) A continuous random variable X follows the probability law

$$f(x) = Ax^2,\ 0 \le x \le 1$$

Determine A and find the probability that (i) X lies between 0·2 and 0·5, (ii) X is less than 0·3, (iii) $1/4 < X < 1/2$ and (iv) $X > 3/4$ given $X > 1/2$.

Ans. $A = 0{\cdot}3$, (i) 0·117, (ii) 0·027, (iii) 15/256 and (iv) 27/56.

(b) If a random variable X has the density function

$$f(x) = \begin{cases} 1/4, & -2 < x < 2 \\ 0, & elsewhere. \end{cases}$$

Obtain (i) $P(X < 1)$, (ii) $P(|X| > 1)$ (iii) $P(2X + 3 > 5)$

(Kerala Univ. B.Sc., Sept.1992)

Hint. (ii) $P(|X| > 1) = P(X > 1 \text{ or } X < -1) = \int_{-2}^{-1} f(x)\,dx + \int_{1}^{2} f(x)\,dx$

or $P(|X| > 1) = 1 - P(|X| \le 1) = 1 - P(-1 \le X \le 1)$

Ans. (i) 3/4, (ii) 1/2 (iii) 1/4.

2. Are any of the following probability mass or density functions? Prove your answer in each case.

(a) $f(x) = x;\ x = \dfrac{1}{16}, \dfrac{3}{16}, \dfrac{1}{4}, \dfrac{1}{2}$

(b) $f(x) = \lambda e^{-\lambda x};\ x \ge 0;\ \lambda > 0$

(c) $f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 4 - 2x, & 1 < x < 2 \\ 0, & elsewhere, \end{cases}$

(Calicut Univ. B. Sc., Oct. 1989)

Ans. (a) and (b) are p.m.f./p.d.f.'s, (c) is not.

3. If f_1 and f_2 are p.d.f.'s and $\theta_1 + \theta_2 = 1$, check if

$$g(x) = \theta_1 f_1(x) + \theta_2 f_2(x), \quad \text{is a p.d.f.}$$

Ans. $g(x)$ is a p.d.f. if $0 \le (\theta_1, \theta_2) \le 1$ and $\theta_1 + \theta_2 = 1$.

4. A continuous random variable X has the probability density function :

$$f(x) = A + Bx\,,\quad 0 \le x \le 1.$$

If the mean of the distribution is $\frac{1}{2}$, find A and B.

Hint : Solve $\int_0^1 f(x)\,dx = 1$ and $\int_0^1 x\, f(x)\,dx = \frac{1}{2}$. Find A and B.

5. For the following density function

$$f(x) = c\,x^2(1 - x)\,,\quad 0 < x < 1\,,$$

find *(i)* the constant c, and *(ii)* mean.

[Calicut Univ. B.Sc.(subs.), 1991]

Ans. *(i)* $c = 12$; *(ii)* mean $= 3/5$.

6. A continuous distribution of a variable X in the range $(-3, 3)$ is defined by

$$\begin{aligned} f(x) &= \tfrac{1}{16}\,(3 + x)^2\,, & -3 \le x \le -1 \\ &= \tfrac{1}{16}\,(6 - 2x^2)\,, & -1 \le x \le 1 \\ &= \tfrac{1}{16}\,(3 - x)^2\,, & 1 \le x \le 3 \end{aligned}$$

(i) Verify that the area under the curve is unity.

(ii) Find the mean and variance of the above distribution.

(Madras Univ. B.Sc., Oct. 1992; Gujarat Univ. B.Sc., Oct. 1986)

Hint: $\int_{-3}^{3} f(x)\,dx = \int_{-3}^{-1} f(x)\,dx + \int_{-1}^{1} f(x)\,dx + \int_{1}^{3} f(x)\,dx$

Ans. Mean=0, Variance=1

7. If the random variable X has the p.d.f.,

$$\begin{aligned} f(x) &= \tfrac{1}{2}\,(x + 1)\,, & -1 < x < 1 \\ &= 0\,, & \text{elsewhere,} \end{aligned}$$

find the coefficient of skewness and kurtosis.

8. *(a)* A random variable X has the probability density function given by

$$f(x) = 6x\,(1 - x)\,,\quad 0 \le x \le 1$$

Find the mean μ, mode and S.D. σ, Compute $P(\mu - 2\sigma < X < \mu + 2\sigma)$.
Find also the mean deviation about the median.

(Lucknow Univ. B.Sc., 1988)

(b) For the continuous distribution

$$dF = y_o\,(x - x^2)\,dx\ ;\quad 0 \le x \le 1\,,\quad y_o \text{ being a constant.}$$

Find *(i)* arithmetic mean, *(ii)* harmonic mean, *(iii)* Median, *(iv)* Mode and *(v)* rth moment about mean. Hence find β_1 and β_2 and show that the distribution is symmetrical. **(Delhi Univ. B.Sc., 1992; Karnatak Univ. B.Sc., 1991)**

Ans. Mean = Median = Mode = $\frac{1}{2}$

(c) Find the mean, mode and median for the distribution,

$$dF(x) = \sin x\, dx, \quad 0 \le x \le \pi/2$$

Ans. 1, $\pi/2$, $\pi/3$

9. If the function $f(x)$ is defined by

$$f(x) = c\, e^{-\alpha x}, \quad 0 \le x < \infty, \quad \alpha > 0$$

(i) Find the value of constant c.

(ii) Evaluate the first four moments about mean.

[Gauhati Univ. B.Sc. 1987]

Ans. *(i)* $c = \alpha$, *(ii)* 0, $1/\alpha^2$, $2/\alpha^3$, $9/\alpha^4$.

10. *(a)* Show that for the exponential distribution

$$dP = y_o \,.\, e^{-x/\sigma}\, dx, \quad 0 \le x < \infty, \quad \sigma > 0$$

the mean and S.D. are both equal to σ and that the interquartile range is $\sigma \log_e 3$. Also find μ_r' and show that $\beta_1 = 4$, $\beta_2 = 9$.

[Agra Univ. B.Sc., 1986 ; Madras Univ. B.Sc., 1987]

(b) Define the harmonic mean (H.M.) of variable X as the reciprocal of the expected value of $1/X$, show that the H.M. of variable which ranges from 0 to ∞ with probability density $\frac{1}{6} x^3 e^{-x}$ is 3.

11. *(a)* Find the mean, variance and the co-efficients β_1, β_2 of the distribution,

$$dF = k\, x^2\, e^{-x}\, dx, \quad 0 < x < \infty.$$

Ans. $k = 1/2$; 3, 3, 4/3 and 5.

(b) Calculate β_1 for the distribution,

$$dF = k\, x\, e^{-x}\, dx, \quad 0 < x < \infty$$

Ans. 2 **[Delhi Univ. B.Sc. (Hons. Subs.), 1988]**

12. A continuous random variable X has a p.d.f. given by

$$f(x) = k\, x\, e^{\lambda x}, \quad x \ge 0, \quad \lambda > 0$$
$$= 0, \text{ otherwise}$$

Determine the constant k, obtain the mean and variance of X.

[Nagpur Univ. B.Sc. 1990]

13. For the probability density function,

$$f(x) = \frac{2(b+x)}{b(a+b)}, \quad -b \le x < 0$$
$$= \frac{2(a-x)}{a(a+b)}, \quad 0 \le x \le a$$

Find mean, median and variance. **[Calcutta Univ. B.Sc. 1984]**

Ans. Mean $= (a-b)/3$, Variance $= (a^2 + b^2 + ab)/18$,
Median $= a - \sqrt{a(a+b)/2}$

(ii) Show that, if terms of order $(a-b)^2/a^2$ are neglected, then

$$\text{mean} - \text{median} = (\text{mean} - \text{mode})/4$$

14. A variable X can assume values only between 0 and 5 and the equation of its frequency curve is

$$y = A \sin \tfrac{1}{5} \pi x, \quad 0 \le x \le 5$$

where A is a constant such that the area under the curve is unity. Determine the value of A and obtain the median and quartiles of the distribution.

Show also that the variance of the distribution is $50\left\{\dfrac{1}{8} - \dfrac{1}{\pi^2}\right\}$.

Ans. 1/10; 2·5, 4/3, 10/3

15. A continuous variable X is distributed over the interval [0, 1] with p.d.f. $ax^2 + bx$, where a, b are constants. If the arithmetic mean of X is 0·5, find the values of a and b.

Ans. $-6, 6$

16. A man leaves his house at the same time every morning and the time taken to journey to work has the following probability density function : less than 30 minutes, zero, between 30 minutes and 60 minutes, uniform with density k ; between 60 minutes and 70 minutes, uniform with density $2k$; and more than 70 minutes, zero. What is the probability that on one particular day he arrives at work later than on the previous day but not more than 5 minutes later.

17. The density function of sheer strength of spot welds is given by

$$\begin{aligned} f(x) &= x/160{,}000 \quad \text{for } 0 \le x \le 400 \\ &= (800 - x)/160{,}000 \quad \text{for } 400 \le x \le 800 \end{aligned}$$

Find the number a such that

Prob. $(X < a) = 0{\cdot}50$ and the number b such that

Prob. $(X < b) = 0{\cdot}90$. Find the mean, median and variance of X.

[Delhi Univ. B.E., 1987]

18. A batch of small calibre ammunition is accepted as satisfactory if none of a sample of five shot falls more than 2 feet from the centre of the target at a given range. If X, the distance from the centre of the target to a given impact point, actually has the density

$$f(x) = k \,.\, 2x e^{-x^2}, \quad 0 < x < 3$$

where k is a number which makes it probability density function, what is the value of k and what is the probability that the batch will be accepted?

[Nagpur Univ. B.E., 1987]

Hint. $\int_0^3 f(x)\, dx = 1 \quad \Rightarrow \quad k = 1/(1 - e^{-9})$

Reqd. Prob. = P [Each of a sample of 5 shots falls within a distance of 2 ft. from the centre]

$$= [P (0 < X < 2)]^5 = \left[\int_0^2 f(x)\, dx \right]^5 = \left[\frac{1 - e^{-4}}{1 - e^{-9}} \right]^5$$

19. A random variable X has the p.d.f. :

$$f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & otherwise \end{cases}$$

Find *(i)* $P\left(X < \frac{1}{2} \right)$, *(ii)* $P\left(\frac{1}{4} < X < \frac{1}{2} \right)$, *(iii)* $P\left(X > \frac{3}{4} \,\middle|\, X > \frac{1}{2} \right)$, and *(iv)* $P\left(X < \frac{3}{4} \,\middle|\, X > \frac{1}{2} \right)$.

(Gorakhpur Univ. B.Sc., 1988)

Ans. *(i)* 1/4 , *(ii)* 3/16 , *(iii)* $\dfrac{P(X > 3/4)}{P(X > 1/2)} = \dfrac{7/16}{3/4} = \dfrac{7}{12}$; *(iv)* $\dfrac{P(1/2 < X < 3/4)}{P(X > 1/2)}$

5·4·3. Continuous Distribution Function. If X is a continuous random variable with the p.d.f. $f(x)$, then the function

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\, dt\ , \quad -\infty < x < \infty . \qquad ...(5\cdot12)$$

is called the *distribution function* (d.f.) or sometimes the *cumulative distribution function* (c.d.f.) of the random variable X.

Remarks 1. $0 \le F(x) \le 1, \quad -\infty < x < \infty$.

2. From analysis (Riemann integral), we know that

$$F'(x) = \frac{d}{dx} F(x) = f(x) \ge 0 \qquad [\because f(x) \text{ is p.d.f.}]$$

$\Rightarrow$ $F(x)$ is non-decreasing function of x.

3. $$F(-\infty) = \lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} \int_{-\infty}^{x} f(x)\, dx = \int_{-\infty}^{-\infty} f(x)\, dx = 0$$

and $$F(+\infty) = \lim_{x \to \infty} F(x) = \lim_{x \to \infty} \int_{-\infty}^{x} f(x)\, dx = \int_{-\infty}^{\infty} f(x)\, dx = 1$$

4. $F(x)$ is a continuous function of x on the right.

5. The discontinuities of $F(x)$ are at the most countable.

6. It may be noted that

$$P(a \le X \le b) = \int_a^b f(x)\, dx = \int_{-\infty}^{b} f(x)\, dx - \int_{-\infty}^{a} f(x)\, dx$$

$$= P(X \le b) - P(X \le a) = F(b) - F(a)$$

Similarly

$$P(a < X < b) = P(a < X \le b) = P(a \le X < b) = \int_a^b f(t)\, dt$$

7. Since $F'(x) = f(x)$, we have

$$\frac{d}{dx} F(x) = f(x) \quad \Rightarrow \quad dF(x) = f(x)\,dx$$

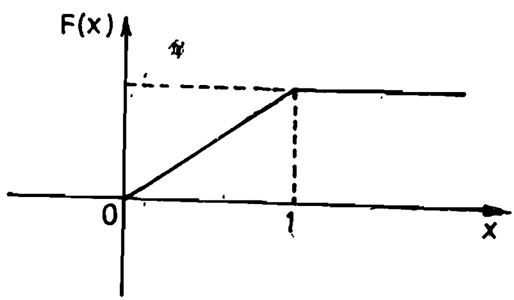
This is known as probability differential of X.

Remarks. 1. It may be pointed out that the properties (2), (3) and (4) above uniquely characterise the distribution functions. This means that any function $F(x)$ satisfying (2) to (4) is the distribution function of some random variable, and any function $F(x)$ violating any one or more of these three properties cannot be the distribution function of any random variable.

2. Often, one can obtain a p.d.f. from a distribution function $F(x)$ by differentiating $F(x)$, provided the derivative exists. For example, consider

$$F_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ x, & \text{for } 0 \le x \le \\ 1, & \text{for } x > 1 \end{cases}$$

The graph of $F(x)$ is given by bold lines. Obviously we see that $F(x)$ is continuous from right as stipulated in (4) and we also see that $F(x)$ is not continuous at $x = 0$ and $x = 1$ and hence is not derivable at $x = 0$ and $x = 1$.



Differentiating $F(x)$ w.r.t. x, we get

$$\frac{d}{dx} F(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \textit{otherwise} \end{cases}$$

[Note the strict inequality in $0 < x < 1$, since $F(x)$ is not derivable at $x = 0$ and $x = 1$]

Let us define

$$f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \textit{otherwise} \end{cases}$$

Then $f(x)$ is a p.d.f. for F.

Example 5·15. *Verify that the following is a distribution function:*

$$F(x) = \begin{cases} 0, & x < -a \\ \frac{1}{2}\left(\frac{x}{a} + 1\right), & -a \le x \le a \\ 1, & x > a \end{cases}$$

(Madras Univ. B.Sc., 1992)

Solution. Obviously the properties *(i), (ii), (iii)* and *(iv)* are satisfied. Also we observe that $F(x)$ is continuous at $x = a$ and $x = -a$, as well.

Now

$$\frac{d}{dx} F(x) = \begin{cases} \frac{1}{2a}, & -a \le x \le a \\ 0, & \text{otherwise} \end{cases}$$

$$= f(x), \text{ say}$$

In order that $F(x)$ is a distribution function, $f(x)$ must be a p.d.f. Thus we have to show that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

Now $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = \int_{-a}^{a} f(x)\, dx = \frac{1}{2a}\int_{-a}^{a} 1 \,.\, dx = 1$

Hence $F(x)$ is a d.f.

Example 5·16. *Suppose the life in hours of a certain kind of radio tube has the probability density function :*

$$f(x) = \frac{100}{x^2}, \quad \text{when } x \ge 100$$

$$= \quad 0, \quad \text{when } x < 100$$

Find the distribution function of the distribution. What is the probabilty that none of three such tubes in a given radio set will have to be replaced during the first 150 hours of operation ? What is the probability that all three of the original tubes will have been replaced during the first 150 hours ? **(Delhi Univ. B.Sc, Oct. 1988)**

Solution. Probability that a tube will last for first 150 hours is given by

$$P(X \le 150) = P(0 < X < 100) + P(100 \le X \le 150)$$

$$= \int_{100}^{150} f(x)\, dx = \int_{100}^{150} \frac{100}{x^2}\, dx = \frac{1}{3}$$

Hence the probability that none of the three tubes will have to be replaced during the first 150 hours is $(1/3)^3 = 1/27$.

The probability that a tube will not last for the first 150 hours is $1 - \frac{1}{3} = \frac{2}{3}$.

Hence the probability that all three of the original tubes will have to be replaced during the first 150 hours is $(2/3)^3 = 8/27$.

Example 5·17. *Suppose that the time in minutes that a person has to wait at a certain station for a train is found to be a random phenomenon, a probability function specified by the distribution function,*

$$\begin{aligned} F(x) &= 0, && \text{for } x \le 0 \\ &= \frac{x}{2}, && \text{for } 0 \le x < 1 \\ &= \frac{1}{2}, && \text{for } 1 \le x < 2 \\ &= \frac{x}{4}, && \text{for } 2 \le x < 4 \\ &= 1, && \text{for } x \ge 4 \end{aligned}$$

(a) Is the Distribution Function continuous? If so, give the formula for its probability density function?

(b) What is the probability that a person will have to wait (i) more than 3 minutes, (ii) less than 3 minutes, and (iii) between 1 and 3 minutes?

(c) What is the conditional probability that the person will have to wait for a train for (i) more than 3 minutes, given that it is more than 1 minute, (ii) less than 3 minutes given that it is more than 1 minute? **(Calicut Univ. B.Sc., 1985)**

Solution. *(a)* Since the value of the distribution function is the same at the points $x = 0, x = 1, x = 2,$ and $x = 4$ given by the two forms of $F(x)$ for $x < 0$ and $0 \le x < 1$, $0 \le x < 1$ and $1 \le x < 2$, $1 \le x < 2$ and $2 \le x < 4$, $2 \le x < 4$ and $x \ge 4$, the distribution function is continuous.

Probability density function $= f(x) = \dfrac{d}{dx} F(x)$

$$\therefore \quad \begin{aligned} f(x) &= 0, && \text{for } x < 0 \\ &= \frac{1}{2}, && \text{for } 0 \le x < 1 \\ &= 0, && \text{for } 1 \le x < 2 \\ &= \frac{1}{4}, && \text{for } 2 \le x < 4 \\ &= 0, && \text{for } x \ge 4 \end{aligned}$$

(b) Let the random variable X represent the waiting time in minutes. Then

(i) Required probability $= P(X > 3) = 1 - P(X \le 3) = 1 - F(3)$

$$= 1 - \frac{1}{4} \cdot 3 = \frac{1}{4}$$

(ii) Required probability $= P(X < 3) = P(X \le 3) - P(X = 3)$

$$= F(3) = \frac{3}{4}$$

(Since, the probability that a continuous variable takes a fixed value is zero) *(iii)* Required Probability $= P(1 < X < 3) = P(1 < X \leq 3)$

$$= F(3) - F(1) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$$

(c) Let A denote the event that a person has to wait for more than 3 minutes and B the event that he has to wait for more than 1 minute. Then

$$P(A) = P(X > 3) = \frac{1}{4} \qquad [\textit{c.f. (b), (i)}]$$

$$P(B) = P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - \frac{1}{2} = \frac{1}{2}$$

$$P(A \cap B) = P(X > 3 \cap X > 1) = P(X > 3) = \frac{1}{4}$$

(i) Required probability is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{1/2} = \frac{1}{2}$$

(ii) Required probability $= P(\bar{A} \mid B) = \dfrac{P(\bar{A} \cap B)}{P(B)}$

Now $P(\bar{A} \cap B) = P(X \leq 3 \cap X > 1) = P(1 < X \leq 3) = F(3) - F(1) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$

$$\therefore \quad P(\bar{A} \mid B) = \frac{1/4}{1/2} = \frac{1}{2}$$

Example 5·18. *A petrol pump is supplied with petrol once a day. If its daily volume X of sales in thousands of litres is distributed by*

$$f(x) = 5(1 - x)^4, \quad 0 \leq x \leq 1,$$

what must be the capacity of its tank in order that the probability that its supply will be exhausted in a given day shall be 0·01 ? **(Madras Univ. B.E., 1986)**

Solution. Let the capacity of the tank (in '000 of litres) be 'a' such that

$$P(X \geq a) = 0{\cdot}01 \quad \Rightarrow \quad \int_a^1 f(x)\,dx = 0{\cdot}01$$

$$\Rightarrow \quad \int_a^1 5(1 - x)^4\,dx = 0{\cdot}01 \quad \text{or} \quad \left[5 \cdot \frac{(1 - x)^5}{(-5)}\right]_a^1 = 0{\cdot}01$$

$$\Rightarrow \quad (1 - a)^5 = 1/100 \quad \text{or} \quad 1 - a = (1/100)^{1/5}$$

$$\therefore \quad a = 1 - (1/100)^{1/5} = 1 - 0{\cdot}3981 = 0{\cdot}6019$$

Hence the capacity of the tank $= 0{\cdot}6019 \times 1000$ *litres* $= 601{\cdot}9$ *litres.*

Example 5·19. *Prove that mean deviation is least when measured from the median.* **[Delhi Univ. B.Sc. (Maths. Hons.), 1989]**

Solution. If $f(x)$ is the probability function of a random variable X, $a \leq X \leq b$, then mean deviation $M(A)$, say, about the point $x = A$ is given by

$$M(A) = \int_a^b |x - A|\, f(x)\,dx$$

$$= \int_a^A |x - A| f(x)\,dx + \int_A^b |x - A| f(x)\,dx$$

$$= \int_a^A (A - x) f(x)\,dx + \int_A^b (x - A) f(x)\,dx \qquad \ldots (1)$$

We want to find the value of 'A' so that $M(A)$ is minimum. From the principle of maximum and minimum in differential calculus, $M(A)$ will be minimum for variations in A if

$$\frac{\partial M(A)}{\partial A} = 0 \quad \text{and} \quad \frac{\partial^2 M(A)}{\partial A^2} > 0 \qquad \ldots (2)$$

Differentiating (1) w.r.t. 'A' under the integral sign, since the functions $(A - x) f(x)$ and $(x - A) f(x)$ vanish at the point $x = A^*$, we get

$$\frac{\partial M(A)}{\partial A} = \int_a^A f(x)\,dx - \int_A^b f(x)\,dx \qquad \ldots (3)$$

Also $\quad \dfrac{\partial M(A)}{\partial A} = \displaystyle\int_a^A f(x)\,dx - \left[1 - \int_a^A f(x)\,dx\right],$

$$\left[\because \int_a^b f(x)\,dx = 1\right]$$

$$= 2\int_a^A f(x)\,dx - 1 = 2F(A) - 1,$$

where $F(\cdot)$ is the distribution function of X. Differentiating again w.r.t. A, we get

$$\frac{\partial^2}{\partial A^2} M(A) = 2f(A) \qquad \ldots (4)$$

Now $\dfrac{\partial M(A)}{\partial A} = 0$, on using (3) gives

$$\int_a^A f(x)\,dx = \int_A^b f(x)\,dx$$

i.e., A is the median value.

Also from (4), we see that

$$\frac{\partial^2 M(A)}{\partial A^2} > 0,$$

assuming that $f(x)$ does not vanish at the median value. Thus mean deviation is least when taken from median.

---

*If $f(x, \theta)$ is a continuous function of both variables x and θ, possessing continuous partial derivatives $\dfrac{\partial^2 f}{\partial x \partial \theta}$, $\dfrac{\partial^2 f}{\partial \theta \partial x}$ and a and b are differentiable functions of θ, then

$$\frac{\partial}{\partial \theta}\left[\int_a^b f(x, \theta)\,dx\right] = \int_a^b \frac{\partial f}{\partial \theta}\,dx + f(b, \theta)\frac{db}{d\theta} - f(a, \theta)\frac{da}{d\theta}$$

EXERCISE 5 (d)

1. *(a)* Explain the terms *(i)* probability differential, *(ii)* probability density function, and *(iii)* distribution function.

(b) Explain what is meant by a random variable. Distinguish between a discrete and a continuous random variable. Define distribution function of a random variable and show that it is monotonic non-decreasing everywhere and continuous on the right at every point.

[Madras Univ. B.Sc. (Stat Main), 1987]

(c) Show that the distribution function $F(x)$ of a random variable X is a non-decreasing function of x. Determine the jump of $F(x)$ at a point x_0 of its discontinuity in terms of the probability that the random variable has the value x_0.

[Calcutta Univ. B.Sc. (Hons.), 1984]

2. The length (in hours) X of a certain type of light bulb may be supposed to be a continuous random variable with probability density function :

$$f(x) = \frac{a}{x^3},\ 1500 < x < 2500$$
$$= 0,\ \text{elsewhere.}$$

Determine the constant a, the distribution function of X, and compute the probability of the event $1,700 \leq X \leq 1,900$.

Ans. $a = 70,31,250$; $F(x) = \frac{a}{2}\left(\frac{1}{22,50,000} - \frac{1}{x^2}\right)$ and

$$P(1,700 < X < 1,900) = F(1,900) - F(1,700) = \frac{a}{2}\left(\frac{1}{28,90,000} - \frac{1}{36,10,000}\right)$$

3. Define the "distribution function" (or cumulative distribution function) of a random variable and state its essential properties.

Show that, whatever the distribution function $F(x)$ of a random variable X, $P[a \leq F(x) \leq b] = b - a,\ 0 \leq a,\ b \leq 1$.

4. *(a)* The distribution function of a random variable X is given by

$$F(x) = \begin{cases} 1 - (1 + x)e^{-x}, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases}$$

Find the corresponding density function of random variable X.

(b) Consider the distribution for X defined by

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ 1 - \frac{1}{4}e^{-x} & \text{for } x \geq 0 \end{cases}$$

Determine $P(x = 0)$ and $P(x > 0)$.

[Allahabad Univ. B.Sc., 1992]

5. *(a)* Let X be a continuous random variable with probability density function given by

$$f(x)=\begin{cases} ax, & 0\le x\le 1 \\ a, & 1\le x\le 2 \\ -ax+3a, & 2\le x\le 3 \\ 0, & elsewhere \end{cases}$$

(i) Determine the constant a.

(ii) Determine $F(x)$, and sketch its graph.

(iii) If three independent observations are made, what is the probability that exactly one of these three numbers is larger than $1{\cdot}5$?

[Rajasthan Univ. M.Sc., 1987]

Ans. *(i)* 1/2, *(iii)* 3/8.

(b) For the density $f_X(x) = k\,e^{-ax}(1-e^{-ax})\ I_{0,\infty}(x)$, find the normalising constant k, $f_X(x)$ and evaluate $P(X>1)$.

[Delhi Univ. B.Sc. (Maths Hons.), 1989]

Ans. $k = 2a$; $F(x) = 1-2e^{-ax}+e^{-2ax}$; $P(X>1) = 2e^{-a}-e^{-2a}$

6. A random variable X has the density function :

$$f(x) = K\cdot\frac{1}{1+x^2},\quad \text{if } -\infty<x<\infty$$
$$= 0,\quad otherwise$$

Determine K and the distribution function.

Evaluate the probability $P(X\ge 0)$. Find also the mean and variance of X.

[Karnatak Univ. B.Sc. 1985]

Ans. $K=1$, $F(x)=\dfrac{1}{\pi}\left\{\tan^{-1}x+\dfrac{\pi}{2}\right\}$, $P(x\ge 0)=1/2$, Mean $=0$, Variance does not exist.

7. A continuous random variable X has the distribution function

$$F(x)=\begin{cases} 0, & \text{if } x\le 1 \\ k(x-1)^4, & \text{if } 1<x\le 3 \\ 1, & \text{if } x>3 \end{cases}$$

Find *(i)* k, *(ii)* the probability density function $f(x)$, and *(iii)* the mean and the median of X.

Ans. *(i)* $k=\dfrac{1}{16}$, *(ii)* $f(x)=\dfrac{1}{4}(x-1)^3$, $1\le x\le 3$

8. Given $f(x)=\begin{cases} kx(1-x), & \text{for } 0<x<1 \\ 0, & elsewhere \end{cases}$

Show that

(i) $k=1/5$, *(ii)* $F(x)=0$ for $x\le 0$ and $F(x)=1-e^{-x/5}$, for $x>0$

Using $F(x)$, show that

(iii) $P(3 < X < 5) = 0.1809$, *(iv)* $P(X < 4) = 0.5507$, *(v)* $P(X > 6) = 0.3012$

9. A bombing plane carrying three bombs flies directly above a railroad track. If a bomb falls within 40 feet of track, the track will be sufficiently damaged to disrupt the traffic. Within a certain bomb site the points of impact of a bomb have the probability density function :

$$\begin{aligned} f(x) &= (100 + x)/10{,}000\,, \quad \textit{when } -100 \le x \le 0 \\ &= (100 - x)/10{,}000\,, \quad \textit{when } 0 \le x \le 100 \\ &= 0\,, \quad \textit{elsewhere} \end{aligned}$$

where x represents the vertical deviation (in feet) from the aiming point, which is the track in this case. Find the distribution function. If all the bombs are used, what is the probability that track will be damaged ?

Hint. Probability that track will be damaged by the bomb is given by

$$\begin{aligned} P(|X| < 40) &= P(-40 < X < 40) \\ &= \int_{-40}^{0} f(x)\,dx + \int_{0}^{40} f(x)\,dx \\ &= \int_{-40}^{0} \frac{100+x}{10{,}000}\,dx + \int_{0}^{40} \frac{100-x}{10{,}000}\,dx = \frac{16}{25} \end{aligned}$$

$\therefore$ Probability that a bomb will not damage the track $= 1 - \frac{16}{25} = \frac{9}{25}$

Probability that none of the three bombs damages the track

$= \left(\frac{9}{25}\right)^3 = 0.046656$

Required probability that the track will be damaged $= 1 - 0.046656 = 0.953344$.

10. The length of time (in minutes) that a certain lady speaks on the telephone is found to be random phenomenon, with a probability function specified by the probability density function $f(x)$ as

$$\begin{aligned} f(x) &= A\,e^{-x/5}\,, \quad \text{for } x \ge 0 \\ &= 0\,, \quad \textit{otherwise} \end{aligned}$$

(a) Find the value of A that makes $f(x)$ a p.d.f.

Ans. $A = 1/5$

(b) What is the probability that the number of minutes that she will talk over the phone is

(i) More than 10 minutes, *(ii)* less than 5 minutes, and *(iii)* between 5 and 10 minutes ? **[Shivaji Univ. B.Sc., 1990]**

Ans. *(i)* $\frac{1}{e^2}$, *(ii)* $\frac{e-1}{e}$, *(iii)* $\frac{e-1}{e^2}$.

11. The probability that a person will die in the time interval (t_1, t_2) is given by

$$A \int_{t_1}^{t_2} f(t)\, dt \; ;$$

where A is a constant and the function $f(t)$ determined from long records, is

$$f(t) = \begin{cases} t^2 (100 - t)^2, & 0 \le t \le 100 \\ 0, & elsewhere \end{cases}$$

Find the probability that a person will die between the ages 60 and 70 assuming that his age is ≥ 50. **[Calcutta Univ. B.A. (Hons.), 1987]**

5·5. Joint Probability Law. Two random variables X and Y are said to be jointly distributed if they are defined on the same probability space. The sample points consist of 2–tuples. If the joint probability function is denoted by $P_{XY}(x, y)$ then the probability of a certain event E is given by

$$P_{XY}(x, y) = P[(X, Y) \in E] \qquad \text{...(5·13)}$$

(X, Y) is said to belong to E, if in the 2 dimensional space the 2–tuples lie in the Borel set B, representing the event E.

5·5·1. Joint Probability Mass Function and Marginal and Conditional Probability Functions. Let X and Y be random variables on a sample space S with respective image sets $X(S) = \{x_1, x_2, \ldots, x_n\}$ and $Y(S) = \{y_1, y_2, \ldots, y_m\}$. We make the product set

$$X(S) \times Y(S) = \{x_1, x_2, \ldots, x_n\} \times \{y_1, y_2, \ldots, y_m\}$$

into a probability space by defining the probability of the ordered pair (x_i, y_j) to be $P(X = x_i, Y = y_j)$ which we write $p(x_i, y_j)$. The function p on $X(S) \times Y(S)$ defined by

$$p_{ij} = P(X = x_i \cap Y = y_j) = p(x_i, y_j) \qquad \text{...(5·14)}$$

is called the *joint probability function* of X and Y and is usually represented in the form of the following table :

X \ Y	y_1	y_2	y_3	...	y_j	...	y_m	Total
x_1	p_{11}	p_{12}	p_{13}	...	p_{1j}	...	p_{1m}	$p_{1\cdot}$
x_2	p_{21}	p_{22}	p_{23}	...	p_{2j}	...	p_{2m}	$p_{2\cdot}$
x_3	p_{31}	p_{32}	p_{33}	...	p_{3j}	...	p_{3m}	$p_{3\cdot}$
⋮	⋮							⋮
x_i	p_{i1}	p_{i2}	p_{i3}	...	p_{ij}	...	p_{im}	$p_{i\cdot}$
⋮	⋮							⋮
x_n	p_{n1}	p_{n2}	p_{n3}	...	p_{nj}	...	p_{nm}	$p_{n\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	...	$p_{\cdot j}$	...	$p_{\cdot m}$	1

$$\therefore \quad \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$$

Suppose the joint distribution of two random variables X and Y is given, then the probability distribution of X is determined as follows :

$$p_X(x_i) = P(X = x_i) = P[X = x_i \cap Y = y_1] + P[X = x_i \cap Y = y_2] + \ldots$$
$$+ P[X = x_i \cap Y = y_j] + \ldots + P[X = x_i \cap Y = y_m]$$
$$= p_{i1} + p_{i2} + \ldots + p_{ij} + \ldots + p_{im}$$
$$= \sum_{j=1}^{m} p_{ij} = \sum_{j=1}^{m} p(x_i, y_j) = p_{i.} \qquad \text{.. (5·14 } a\text{)}$$

and is known as *marginal probability function of X.*

Also $\sum_{i=1}^{n} p_{i.} = p_{1.} + p_{2.} + \ldots + p_{n.} = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = 1$

Similarly, we can prove that

$$p_Y(y_j) = P(Y = y_j) = \sum_{i=1}^{n} p_{ij} = \sum_{i=1}^{n} p(x_i, y_j) = p_{.j} \qquad \text{... (5·14 } b\text{)}$$

which is the *marginal probability function of Y.*

Also

$$P[X = x_i \mid Y = y_j] = \frac{P[X = x_i \cap Y = y_j]}{P[Y = y_j]} = \frac{p(x_i, y_j)}{p(y_j)} = \frac{p_{ij}}{p_{.j}}$$

This is known as *conditional probability function of X given Y =* y_j

Similarly

$$P[Y = y_j \mid X = x_i] = \frac{p(x_i, y_j)}{p(x_i)} = \frac{p_{ij}}{p_{i.}} \qquad \text{... (5·14 } c\text{)}$$

is the *conditional probability function of Y given X =* x_i

Also $\sum_{i=1}^{n} \frac{p_{ij}}{p_{.j}} = \frac{p_{1j} + p_{2j} + \ldots + p_{ij} + \ldots + p_{nj}}{p_{.j}} = \frac{p_{.j}}{p_{.j}} = 1$

Similarly

$$\sum_{j=1}^{n} \frac{p_{ij}}{p_{i.}} = 1$$

Two random variables X and Y are said to be *independent* if

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j), \qquad \text{... (5·14 } d\text{)}$$

otherwise they are said to be dependent.

5·5·2. Joint Probability Distribution Function. Let (X, Y) be a two-dimensional random variable then their joint distribution function is denoted by $F_{XY}(x, y)$ and it represents the probability that simultaneously the observation

(X, Y) will have the property ($X \leq x$ and $Y \leq y$), *i.e.*,

$$F_{XY}(x, y) = P(-\infty < X \leq x, -\infty < Y \leq y)$$

$$= \int_{-\infty}^{x} \left[\int_{-\infty}^{y} f_{XY}(x, y)\, dx\, dy \right. \qquad \ldots (5\cdot15)$$

(For continuous variables)

where $\qquad f_{XY}(x, y) \geq 0$

And $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x, y)\; dx\, dy = 1 \;\; or \;\; \sum_x \sum_y f(x, y) = 1$

Properties of Joint Distribution Function

1. (*i*) For the real numbers a_1, b_1, a_2 and b_2

$$P(a_1 < X \leq b_1, a_2 < Y \leq b_2) = F_{XY}(b_1, b_2) + F_{XY}(a_1, a_2) - F_{XY}(a_1, b_2) - F_{XY}(b_1, a_2)$$

[For proof, See Example 5·29]

(*ii*) Let $a_1 < a_2$, $b_1 < b_2$. We have

$$(X \leq a_1, Y \leq a_2) + (a_1 < X \leq b_1, Y \leq a_2) = (X \leq b_1, Y \leq a_2)$$

and the events on the L.H.S. are mutually exclusive. Taking probabilities on both-sides, we get :

$$F(a_1, a_2) + P(a_1 < X \leq b_1, Y \leq a_2) = F(b_1, a_2)$$

$$\Rightarrow \qquad F(b_1, a_2) - F(a_1, a_2) = P(a_1 < X \leq b_1, Y \leq a_2)$$

$$\therefore \qquad F(b_1, a_2) \geq F(a_1, a_2) \qquad [\text{Since } P(a_1 < X \leq b_1, Y \leq a_2) \geq 0]$$

Similarly it follows that

$$F(a_1, b_2) - F(a_1, a_2) = P(X \leq a_1, a_2 < Y \leq b_2) \geq 0$$

$$\Rightarrow \qquad F(a_1, b_2) \geq F(a_1, a_2),$$

which shows that $F(x, y)$ is monotonic non-decreasing function.

2. $\quad F(-\infty, y) = 0 = F(x, -\infty),\; F(+\infty, +\infty) = 1$

3. If the density function $f(x, y)$ is continuous at (x, y) then

$$\frac{\partial^2 F}{\partial x\, \partial y} = f(x, y)$$

5·5·3. Marginal Distribution Functions. From the knowledge of joint distribution function $F_{XY}(x, y)$, it is possible to obtain the individual distribution functions, $F_X(x)$ and $F_Y(y)$ which are termed as marginal distribution function of X and Y respectively with respect to the joint distribution function $F_{XY}(x, y)$.

$$F_X(x) = P(X \leq x) = P(X \leq x, Y < \infty) = \lim_{y \to \infty} F_{XY}(x, y).$$

$$= F_{XY}(x, \infty) \qquad \ldots (5\cdot16)$$

Similarly, $F_Y(y) = P(Y \leq y) = P(X < \infty, Y \leq y)$

$$= \lim_{x \to \infty} F_{XY}(x, y) = F_{XY}(\infty, y)$$

$F_X(x)$ is termed as the marginal distribution function of X corresponding to the joint distribution function $F_{XY}(x, y)$ and similarly $F_Y(y)$ is called marginal distribution function of the random variable Y corresponding to the joint distribution function $F_{XY}(x, y)$.

In the case of jointly discrete random variables, the marginal distribution functions are given as

$$F_X(x) = \sum_y P(X \le x, Y = y),$$

$$F_Y(y) = \sum_x P(X = x, Y \le y)$$

Similarly in the case of jointly continuous random variable, the marginal distribution functions are given as

$$F_X(x) = \int_{-\infty}^{x} \left\{ \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \right\} dx$$

$$F_Y(y) = \int_{-\infty}^{y} \left\{ \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \right\} dy$$

5·5·4 Joint Density Function, Marginal Density Functions. From the joint distribution function $F_{XY}(x, y)$ of two dimensional continuous random variable we get the joint probabilty density function by differentiation as follows :

$$f_{XY}(x, y) = \partial^2 F(x, y)/\partial x \partial y$$

$$= \lim_{\delta x \to 0,\, \delta y \to 0} \frac{P(x \le X \le x + \delta x, y \le Y \le y + \delta y)}{\delta x\ \delta y}$$

Or it may be expressed in the following way also :

"The probability that the point (x, y) will lie in the infinitesimal rectangular region, of area $dx\, dy$ is given by

$$P\left\{ x - \tfrac{1}{2}dx \le X \le x + \tfrac{1}{2}dx, y - \tfrac{1}{2}dy \le Y \le y + \tfrac{1}{2}dy \right\} = dF_{XY}(x, y)$$

and is denoted by $f_{XY}(x, y)\, dx\, dy$, where the function $f_{XY}(x, y)$ is called the joint probability density function of X and Y.

The marginal probability function of Y and X are given respectively

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \qquad \text{(for continuous variables)}$$

$$= \sum_x p_{XY}(x, y) \qquad \text{(for discrete variables)}$$

...(5·17)

and $$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \qquad \text{(for continuous variables)}$$

$$= \sum_y p_{XY}(x, y) \qquad \text{(for discrete variables)} \qquad (5{\cdot}17a)$$

The marginal density functions of X and Y can be obtained in the following manner also.

$$\left.\begin{aligned} f_X(x) &= \frac{dF_X(x)}{dx} = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \\ \text{and} \quad f_Y(y) &= \frac{dF_Y(y)}{dy} = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \end{aligned}\right\} \quad \ldots (5{\cdot}17\, b)$$

Important Remark. If we know the joint p.d.f. (p.m.f.) $f_{XY}(x, y)$ of two random variables X and Y, we can obtain the individual distributions of X and Y in the form of their marginal p.d.f.'s (p.m.f's) $f_X(x)$ and $f_Y(y)$ by using (5·17) and (5·17a). However, the converse is not true *i.e., from the marginal distributions of two jointly distributed random variables, we cannot determine the joint distributions of these two random variables.*

To verify this, it will suffice to show that two different joint p.m.f's (p.d.f.'s) have the same marginal distribution for X and the same marginal distribution for Y. We give below two joint discrete probability distributions which have the same marginal distributions.

JOINT DISTRIBUTIONS HAVING SAME MARGINALS

Probability Distribution I

$Y \backslash X$	0	1	$f_Y(y)$
0	0·28	0·37	0·65
1	0·22	0·13	0·35
$f_X(x)$	0·50	0·50	1·00

Probability Distribution II

$Y \backslash X$	0	1	$f_Y(y)$
0	0·35	0·30	0·65
1	0·15	0·20	0·35
$f_X(x)$	0·50	0·50	1·00

As an illustration for continuous random variables, let (X, Y) be continuous r.v. with joint p.d.f.

$$f_{XY}(x, y) = x + y \;;\quad 0 \le (x, y) \le 1 \qquad \ldots (5{\cdot}17\, c)$$

The marginal p.d.f.'s of X and Y are given by :

$$f_X(x) = \int_0^1 f(x, y)\, dy = \int_0^1 (x + y)\, dy = \left| xy + \frac{y^2}{2} \right|_0^1$$

$$\left.\begin{aligned} \Rightarrow \qquad & f_X(x) = x + \tfrac{1}{2} \ ; && 0 \le x \le 1 \\ \text{Similarly} \quad & f_Y(y) = \int_0^1 f(x, y)\, dx = y + \tfrac{1}{2} \ ; && 0 \le y \le 1 \end{aligned}\right\} \qquad \dots (5\cdot17\, d)$$

Consider another continuous joint p.d.f.

$$g(x, y) = \left(x + \tfrac{1}{2}\right)\left(y + \tfrac{1}{2}\right) \ ; \quad 0 \le (x, y) \le 1 \qquad \dots (5\cdot17\, e)$$

Then marginal p.d.f.'s of X and Y are given by :

$$g_1(x) = \int_0^1 g(x, y)\, dy = \left(x + \tfrac{1}{2}\right) \int_0^1 \left(y + \tfrac{1}{2}\right) dy$$

$$= \left(x + \tfrac{1}{2}\right) \left| \frac{y^2}{2} + \frac{1}{2} y \right|_0^1$$

$$\left.\begin{aligned} \Rightarrow \qquad & g_1(x) = x + \tfrac{1}{2} \ ; && 0 \le x \le 1 \\ \text{Similarly} \quad & g_2(y) = y + \tfrac{1}{2} \ ; && 0 \le y \le 1 \end{aligned}\right\} \qquad \dots (5\cdot17\, f)$$

(5·17 d) and (5·17 f) imply that the two joint p.d.f.'s in (5·17 c) and (5·17 e) have the same marginal p.d.f.'s (5·17 d) or (5·17 f).

Another illustration of continuous r.v.'s is given in Remark to Bivariate Normal Distribution, § 10·10·2.

5·5·5. The Conditional Distribution Function and Conditinal Probability Density Function. For two diamensional random variable (X, Y), the joint distribution function $F_{XY}(x, y)$, for any real numbers x and y is given by

$$F_{XY}(x, y) = P(X \le x, Y \le y)$$

Now let A be the event $(Y \le y)$ such that the event A is said to occur when Y assumes values up to and inclusive of y.

Using conditional probabilities we may now write

$$F_{XY}(x, y) = \int_{-\infty}^{x} P[A \mid X = x]\, dF_X(x) \qquad \dots (5\cdot18)$$

The *conditional distribution function* $F_{Y|X}(y \mid x)$ denotes the distribution function of Y when X has already assumed the particular value x. Hence

$$F_{Y|X}(y \mid x) = P[Y \le y \mid X = x] = P[A \mid X = x]$$

Using this expression, the joint distribution function $F_{XY}(x, y)$ may be expressed in terms of the conditional distribution function as follows :

$$F_{XY}(x, y) = \int_{-\infty}^{x} F_{Y|X}(y \mid x)\, dF_X(x) \qquad \dots (5\cdot18\, a)$$

Similarly

$$F_{XY}(x, y) = \int_{-\infty}^{y} F_{X|Y}(x \mid y)\, dF_Y(y) \qquad \dots (5\cdot18\, b)$$

The *conditional probability density function* of Y given X for two random variables X and Y which are jointly continuously distributed is defined as follows, for two real numbers x and y:

$$f_{Y|X}(y|x) = \frac{\partial}{\partial y} F_{Y|X}(y|x) \qquad \dots (5\cdot19)$$

Remarks : 1. $f_X(x) > 0$, then

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

Proof. We have

$$F_{XY}(x, y) = \int_{-\infty}^{x} F_{Y|X}(y|x)\, dF_X(x)$$

$$= \int_{-\infty}^{x} F_{Y|X}(y|x)\, f_X(x)\, dx$$

Differentiating w.r.t. x, we get

$$\frac{\partial}{\partial x} F_{XY}(x, y) = F_{Y|X}(y|x)\, f_X(x)$$

Differentiating w.r.t. y, we get

$$\frac{\partial}{\partial y}\left[\frac{\partial}{\partial x} F_{XY}(x, y)\right] = f_{Y|X}(y|x)\, f_X(x)$$

$$\Rightarrow \qquad f_{XY}(x, y) = f_{Y|X}(y|x)\, f_X(x)$$

$$\Rightarrow \qquad f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

2. If $f_Y(y) > 0$, then

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

3. In terms of the differentials, we have

$$P(x < X \leq x + dx \mid y < Y \leq y + dy)$$

$$= \frac{P(x < X \leq x + dx,\ y < Y \leq y + dy)}{P(y < Y \leq y + dy)}$$

$$= \frac{f_{XY}(x, y)\, dx\, dy}{f_Y(y)\, dy} = f_{X|Y}(x|y)\, dx$$

Whence $f_{X|Y}(x|y)$ may be interpreted as the conditional density function of X on the assumption $Y = y$.

5·5·6. Stochastic Independence. Let us consider two random variables X and Y (of discrete or continuous type) with joint p.d.f. $f_{XY}(x, y)$ and marginal p.d.f.'s $f_X(x)$ and $g_Y(y)$ respectively. Then by the compound probability theorem

$$f_{XY}(x, y) = f_X(x)\, g_Y(y|x)$$

where $g_Y(y \mid x)$ is the conditional p.d.f. of Y for given value of $X = x$.

If we assume that $g(y \mid x)$ does not depend on x, then by the definition of marginal p.d.f.'s, we get for continuous r.v.'s

$$g(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

$$= \int_{-\infty}^{\infty} f_X(x)\, g(y \mid x)\, dx$$

$$= g(y \mid x) \int_{-\infty}^{\infty} f_X(x)\, dx$$

[since $g(y \mid x)$ does not depend on x]

$$= g(y \mid x) \qquad [\because f(.) \text{ is p.d.f. of } X]$$

Hence

$$g(y) = g(y \mid x)$$

and $$f_{XY}(x, y) = f_X(x)\, g_Y(y) \qquad \ldots (*)$$

provided $g(y \mid x)$ does not depend on x. This motivates the following definition of independent random variables.

Independent Random variables. *Two r.v.'s X and Y with joint p.d.f. $f_{XY}(x, y)$ and marginal p.d.f.'s $f_X(x)$ and $g_Y(y)$ respectively are said to be stochastically independent if and only if*

$$f_{X,Y}(x, y) = f_X(x)\, g_Y(y) \qquad \ldots (5\cdot20)$$

Remarks. 1. In terms of the distribution function, we have the following definition:

Two jointly distributed random variables X and Y are stochastically independent if and only if their joint distribution function $F_{X,Y}(\,.\,,\,.\,)$ is the product of their marginal distribution functions $F_X(\cdot)$ and $G_Y(\cdot)$, i.e., if for real (x, y)

$$F_{X,Y}(x, y) = F_X(x)\, G_Y(y) \qquad \ldots (5\cdot20\, a)$$

2. The variables which are not stochastically independent are said to be stochastically dependent.

Theorem 5·8. *Two random variables X and Y with joint p.d.f. $f(x, y)$ are stochastically independent if and only if $f_{X,Y}(x, y)$ can be expressed as the product of a non-negative function of x alone and a non-negative function of y alone, i.e., if*

$$f_{X,Y}(x, y) = h_X(x)\, k_Y(y) \qquad \ldots (5\cdot21)$$

where $h(\cdot) \geq 0$ and $k(\cdot) \geq 0$.

Proof. If X and Y are independent then by definition, we have

$$f_{X,Y}(x, y) = f_X(x) \cdot g_Y(y)$$

where $f(x)$ and $g(y)$ are marginal p.d.f. of X and Y respectively. Thus condition (5·21) is satisfied.

Conversely if (5·21) holds, then we have to prove that X and Y are independent. For continuous random variables X and Y, the marginal p.d.f.'s are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{-\infty}^{\infty} h(x)\, k(y)\, dy$$

$$= h(x) \int_{-\infty}^{\infty} k(y)\, dy = c_1\, h(x), \text{ say} \qquad \ldots (*)$$

and $$g_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_{-\infty}^{\infty} h(x)\, k(y)\, dx$$

$$= k(y) \int_{-\infty}^{\infty} h(x)\, dx = c_2\, k(y), \text{ say} \qquad \ldots (**)$$

where c_1 and c_2 are constants independent of x and y. Moreover

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

$$\Rightarrow \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)\, k(y)\, dx\, dy = 1$$

$$\Rightarrow \quad \left(\int_{-\infty}^{\infty} h(x)\, dx\right)\left(\int_{-\infty}^{\infty} k(y)\, dy\right) = 1$$

$$\Rightarrow \quad c_2\, c_1 = 1 \qquad \ldots (***)$$

Finally, we get

$$f_{X,Y}(x, y) = h_X(x)\, k_Y(y) = c_1 c_2\, h_X(x)\, k_Y(y) \qquad [\text{using } (***)]$$

$$= (c_1 h_X(x))\, (c_2 k_Y(y))$$

$$= f_X(x)\, g_Y(y) \qquad [\text{from } (*) \text{ and } (**)]$$

$\Rightarrow$ X and Y are stochastically independent.

Theorem 5·9. *If the random variables X and Y are stochastically independent, then for all possible selections of the corresponding pairs of real numbers (a_1, b_1), (a_2, b_2) where $a_i \le b_i$ for all $i = 1, 2$ and where the values $\pm\infty$ are allowed, the events $(a_1 < X \le b_1)$ and $(a_2 < Y \le b_2)$ are independent, i.e.,*

$$P[(a_1 < X \le b_1) \cap (a_2 < Y \le b_2)] = P(a_1 < X \le b_1)\, P(a_2 < Y \le b_2)$$

Proof. Since X and Y are stochastically independent, we have in the usual notations

$$f_{X,Y}(x, y) = f_X(x)\, g_Y(y) \qquad \ldots (*)$$

In case of continuous r.v.'s, we have

$$P[(a_1 < X \le b_1) \cap (a_2 < Y \le b_2)] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y)\, dx\, dy$$

$$= \left(\int_{a_1}^{b_1} f_X(x)\ dx \right) \left(\int_{a_2}^{b_2} g_Y(y)\ dy \right) \qquad [\text{ from } (*)]$$

$$= P(a_1 < X \leq b_1)\, P(a_2 < Y \leq b_2)$$

as desired.

Remark. In case of discrete r.v.'s theorems 5·8 and 5·9 can be proved on replacing integration by summation over the given range of the variables.

Example 5.20. *For the following bivariate probability distribution of X and Y, find*

(i) $P(X \leq 1, Y = 2)$, *(ii)* $P(X \leq 1)$, *(iii)* $P(Y = 3)$, *(iv)* $P(Y \leq 3)$ *and* *(v)* $P(X < 3, Y \leq 4)$

X \ Y	1	2	3	4	5	6
0	0	0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{3}{32}$
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
2	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$	0	$\frac{2}{64}$

Solution. The marginal distributions are given below :

X \ Y	1	2	3	4	5	6	$p_X(x)$
0	0	0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{3}{32}$	$\frac{8}{32}$
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{10}{16}$
2	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$	0	$\frac{2}{64}$	$\frac{8}{64}$
$p_Y(y)$	$\frac{3}{32}$	$\frac{3}{32}$	$\frac{11}{64}$	$\frac{13}{64}$	$\frac{6}{32}$	$\frac{16}{64}$	$\Sigma p(x) = 1$ $\Sigma p(y) = 1$

(i) $P(X \leq 1, Y = 2) = P(X = 0, Y = 2) + P(X = 1, Y = 2)$

$$= 0 + \frac{1}{16} = \frac{1}{16}$$

(ii) $P(X \leq 1) = P(X = 0) + P(X = 1)$

$$= \frac{8}{32} + \frac{10}{16} = \frac{7}{8} \qquad (\text{From above table})$$

(iii) $P(Y = 3) = \frac{11}{64}$ (From above table)

(iv) $P(Y \leq 3) = P(Y = 1) + P(Y = 2) + P(Y = 3)$

$$= \frac{3}{32} + \frac{3}{32} + \frac{11}{64} = \frac{23}{64}$$

(v) $P(X<3, Y\le 4) = P(X=0, Y\le 4) + P(X=1, Y\le 4)$

$$+P(X=2, Y\le 4)$$

$$=\left(\frac{1}{32}+\frac{2}{32}\right)+\left(\frac{1}{16}+\frac{1}{16}+\frac{1}{8}+\frac{1}{8}\right)$$

$$+\left(\frac{1}{32}+\frac{1}{32}+\frac{1}{64}+\frac{1}{64}\right)=\frac{9}{16}$$

Example 5·21. *The joint probability distribution of two random variables X and Y is given by :*

$$p(x,y)=\frac{2}{n(n+1)},\quad \begin{array}{l} x=1,2,\ldots,n \\ y=1,2,\ldots,x \end{array}$$

Examine whether X and Y are independent. **(Calicut Univ. B.Sc., 1991)**

Solution. The joint probability distribution table along with the marginal distributions of X and Y is given below.

$Y \backslash X$	1	2	3	....	n	$p_Y(y)$
1	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$	....	$\frac{2}{n(n+1)}$	$\frac{2n}{n(n+1)}$
2	–	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$	....	$\frac{2}{n(n+1)}$	$\frac{2(n-1)}{n(n+1)}$
3	–	–	$\frac{2}{n(n+1)}$	....	$\frac{2}{n(n+1)}$	$\frac{2(n-2)}{n(n+1)}$
⋮	⋮					
$n-1$	–	–	–	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$	$\frac{2\times 2}{n(n+1)}$
n	–	–	–	–	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$
$p_X(x)$	$\frac{2}{n(n+1)}$	$\frac{2\times 2}{n(n+1)}$	$\frac{2\times 3}{n(n+1)}$	....	$\frac{2\times n}{n(n+1)}$	

Note that $y=1,2,\ldots,x$.

When $x=1$, $y=1$; when $x=2$, $y=1,2$; when $x=3$, $y=1,2,3$ and so on.

From the above table, we see that

$$p_{XY}(x,y)\neq p_X(x)\,p_Y(y)\ ;\quad \forall\ x,y$$

$\Rightarrow$ X and Y are not independent.

Example 5·22. *Given the following bivariate probability distribution, obtain (i) marginal distributions of X and Y, (ii) the conditional distribution of X given Y = 2.*

Y \ X	−1	0	1
0	1/15	2/15	1/15
1	3/15	2/15	1/15
2	2/15	1/15	2/15

(Mysore Univ. B.Sc., Oct. 1987)

Solution.

Y \ X	−1	0	1	$\sum_x p(x,y)$
0	1/15	2/15	1/15	4/15
1	3/15	2/15	1/15	6/15
2	2/15	1/15	2/15	5/15
$\sum_y p(x,y)$	6/15	5/15	4/15	1

(i) Marginal distribution of X. From the above table, we get

$$P(X=-1)=\frac{6}{15}=\frac{2}{5}\ ;\quad P(X=0)=\frac{5}{15}=\frac{1}{3}\ ;\quad P(X=1)=\frac{4}{15}$$

Marginal distribution of Y :

$$P(Y=0)=\frac{4}{15}\ ;\quad P(Y=1)=\frac{6}{15}=\frac{2}{5}\ ;\quad P(Y=2)=\frac{5}{15}=\frac{1}{3}$$

(ii) Conditional distribution of X given Y = 2. We have

$$P(X=x\cap Y=2)=P(Y=2).P(X=x\mid Y=2)$$

$$\Rightarrow\qquad P(X=x\mid Y=2)=\frac{P(X=x\cap Y=2)}{P(Y=2)}$$

$$\therefore\qquad P(X=-1\mid Y=2)=\frac{P(X=-1\cap Y=2)}{P(Y=2)}=\frac{2/15}{1/3}=\frac{2}{5}$$

Example 5·23. *X and Y are two random variables having the joint density function, $f(x,y)=\frac{1}{27}(2x+y)$, where x and y can assume only the integer values 0, 1 and 2. Find the conditional distribution of Y for X = x.*

[South Gujarat Univ. B.Sc., 1988]

Solution. The joint probability function

$$f(x,y)=\frac{1}{27}(2x+y);\quad x=0,1,2;\quad y=0,1,2$$

gives the following table of joint probability distribution of X and Y.

JOINT PROBABILITY DISTRIBUTION $f(x,y)$ OF X AND Y

$X\downarrow$ \ $Y\rightarrow$	0	1	2	$f_X(x)$
0	0	1/27	2/27	3/27
1	2/27	3/27	4/27	9/27
2	4/27	5/27	6/27	15/27

For example $f(0,0)=\frac{1}{27}(0+2\times 0)=0$

$f(1,0)=\frac{1}{27}(0+2\times 1)=\frac{2}{27}$; $f(2,0)=\frac{1}{27}(0+2\times 2)=\frac{4}{27}$
and so on.

The marginal probability distribution of X is given by

$$f_X(x)=\sum_y f(x,y),$$

and is tabulated in last column of above table.

The conditional distribution of Y for $X=x$ is given by

$$f_{Y|X}(Y=y\,|\,X=x)=\frac{f(x,y)}{f_X(x)}$$

and is obtained in the following table.

CONDITIONAL DISTRIBUTION OF Y FOR $X=x$

X \ Y	0	1	2
0	0	1/3	2/3
1	2/9	3/9	4/9
2	4/15	5/15	6/15

Example 5·24. *Two discrete random variables X and Y have the joint probability density function :*

$$p(x,y)=\frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!\,(x-y)!},\quad y=0,1,2,\ldots,x\ ;\ x=0,1,2,\ldots$$

where λ, p are constants with $\lambda > 0$ and $0 < p < 1$.

Find (i) The marginal probability density functions of X and Y.

(ii) The conditional distribution of Y for a given X and of X for a given Y.

(Poona Univ. B.Sc., 1986 ; Nagpur Univ. M.Sc., 1989)

Solution. *(i)*

$$p_X(x) = \sum_{y=0}^{x} p(x,y) = \sum_{y=0}^{x} \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!\,(x-y)!}$$

$$= \frac{\lambda^x e^{-\lambda}}{x!} \sum_{y=0}^{x} \frac{x!\,p^y (1-p)^{x-y}}{y!\,(x-y)!} = \frac{\lambda^x e^{-\lambda}}{x!} \sum_{y=0}^{x} {}^xC_y p^y (1-p)^{x-y}$$

$$= \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

which is the probability function of a Poisson distribution with parameter λ.

$$p_Y(y) = \sum_{x=0}^{\infty} p(x,y) = \sum_{x=y}^{\infty} \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!\,(x-y)!}$$

$$= \frac{(\lambda p)^y e^{-\lambda}}{y!} \sum_{x=y}^{\infty} \frac{[\lambda(1-p)]^{x-y}}{(x-y)!} = \frac{(\lambda p)^y e^{-\lambda}}{y!} e^{\lambda(1-p)}$$

$$= \frac{e^{-\lambda p} (\lambda p)^y}{y!}, \quad y = 0, 1, 2, \ldots,$$

which is the probability function of a Poisson distribution with parameter λp.

(ii) The conditional distribution of Y for given X is

$$p_{Y|X}(y \mid x) = \frac{p_{XY}(x,y)}{p_X(x)} = \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y} x!}{y!\,(x-y)!\,\lambda^x e^{-\lambda}}$$

$$= \frac{x!}{y!\,(x-y)!} p^y (1-p)^{x-y} = {}^xC_y\, p^y (1-p)^{x-y}, \quad x > y$$

The conditional probability distribution of X for given Y is

$$p_{X|Y}(x \mid y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

$$= \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!\,(x-y)!} \cdot \frac{y!}{e^{-\lambda p} (\lambda p)^y} \qquad [\,c.f.\ \text{Part (i)}\,]$$

$$= \frac{e^{-\lambda q} (\lambda q)^{x-y}}{(x-y)!}\ ; \quad q = 1-p,\ x > y$$

Example 5·25. *The joint p.d.f. of two random variables X and Y is given by :*

$$f(x,y) = \frac{9(1+x+y)}{2(1+x)^4 (1+y)^4}\ ; \quad \begin{pmatrix} 0 \le x < \infty \\ 0 < y < \infty \end{pmatrix}$$

Find the marginal distributions of X and Y, and the conditional distribution of Y for X = x.

Solution. Marginal p.d.f. of X is given by

$$f_X(x) = \int_0^\infty f(x,y)\, dy$$

$$= \frac{9}{2(1+x)^4} \int_0^\infty \frac{(1+y)+x}{(1+y)^4}\, dy$$

$$= \frac{9}{2(1+x)^4} \cdot \int_0^\infty \left[(1+y)^{-3} + x\,(1+y)^{-4}\right] dy$$

$$= \frac{9}{2(1+x)^4} \left[\left| \frac{-1}{2(1+y)^2} \right|_0^\infty + x \left| \frac{-1}{3(1+y)^3} \right|_0^\infty \right]$$

$$= \frac{9}{2(1+x)^4} \cdot \left[\frac{1}{2} + \frac{x}{3} \right]$$

$$= \frac{3}{4} \cdot \frac{3+2x}{(1+x)^4}\ ;\ 0 < x < \infty$$

Since $f(x,y)$ is symmetric in x and y, the marginal p.d.f. of Y is given by

$$f_Y(y) = \int_0^\infty f(x,y)\, dx$$

$$= \frac{3}{4} \cdot \frac{3+2y}{(1+y)^4}\ ;\ 0 < y < \infty$$

The conditional distribution of Y for $X = x$ is given by

$$f_{XY}(Y = y \mid X = x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

$$= \frac{9(1+x+y)}{2(1+x)^4(1+y)^4} \cdot \frac{4(1+x)^4}{3(3+2x)}$$

$$= \frac{6(1+x+y)}{(1+y)^4(3+2x)}\ ;\ 0 < y < \infty$$

Example 5·26. *The joint probability density function of a two-dimensional random variable (X,Y) is given by*

$$f(x,y) = 2\ ;\ 0 < x < 1,\ 0 < y < x$$
$$= 0,\ \textit{elsewhere}$$

(i) Find the marginal density functions of X and Y,

(ii) find the conditional density function of Y given X = x and conditional density function of X given Y = y, and

(iii) check for independence of X and Y.

[M.S.Baroda Univ. B.Sc., 1987; Karnataka Univ. B.Sc., Oct. 1988]

Solution. Evidently $f(x, y) \geq 0$ and

$$\int_0^1 \int_0^x 2\, dx\, dy = 2 \int_0^1 x\, dx = 1$$

(i) The marginal p.d.f.'s of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy = \int_0^x 2\, dy = 2x, \quad 0 < x < 1$$

$$= 0, \text{ elsewhere}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx = \int_y^1 2dx = 2(1 - y), \quad 0 < y < 1$$

$$= 0, \text{ elsewhere}$$

(ii) The conditional density function of Y given X is

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x}, \quad 0 < x < 1$$

The conditional density function of X given Y is

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{2}{2(1-y)} = \frac{1}{(1-y)}, \quad 0 < y < 1$$

(iii) Since $f_X(x)\, f_Y(y) = 2(2x)(1 - y) \neq f_{XY}(x, y)$, X and Y are not independent.

Example 5·27. *A gun is aimed at a certain point (origin of the coordinate system). Because of the random factors, the actual hit point can be any point (X,Y) in a circle of radius R about the origin. Assume that the joint density of X and Y is constant in this circle given by :*

$$f_{XY}(x, y) = k, \text{ for } x^2 + y^2 \leq R^2$$

$$= 0, \textit{ otherwise}$$

(i) Compute k, (ii) show that

$$f_X(x) = \frac{2}{\pi R} \left\{ 1 - \left(\frac{x}{R} \right)^2 \right\}^{1/2}, \quad \text{for } -R \leq x \leq R$$

$$= 0, \textit{ otherwise}$$

[Calcutta Univ. B.Sc.(Stat. Hons.),1987]

Solution. *(i)* The constant k is computed from the consideration that the total probability is 1, *i.e.*,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1 \quad \Rightarrow \quad \iint_{x^2 + y^2 \leq R^2} k\, dx\, dy = 1$$

$$\Rightarrow \quad 4 \iint_I k\,dx\,dy = 1$$

where region I is the first quadrant of the circle $x^2 + y^2 = R^2$.

$$\Rightarrow \quad 4k \int_0^R \left(\int_0^{\sqrt{R^2 - x^2}} 1 \, . \, dy \right) dx = 1$$

$$\Rightarrow \quad 4k \int_0^R \sqrt{R^2 - x^2} \, dx = 1$$

$$\Rightarrow \quad 4k \left| x \sqrt{R^2 - x^2} + \frac{R^2}{2} \sin^{-1}\left(\frac{x}{R}\right) \right|_0^R = 1$$

$$\Rightarrow \quad 4k \cdot \left(\frac{R^2}{2} \cdot \frac{\pi}{2} \right) = 1 \quad \Rightarrow \quad k = \frac{1}{\pi R^2}$$

$$\therefore \quad f_{XY}(x, y) = 1/(\pi R^2) \; ; \quad x^2 + y^2 \leq R^2$$
$$= 0, \quad otherwise$$

(ii) $$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \frac{1}{\pi R^2} \int_{-\sqrt{R^2 - x^2}}^{\sqrt{R^2 - x^2}} 1 \cdot dy$$

$\left[because \; x^2 + y^2 \leq R^2 \Rightarrow -\left(R^2 - x^2\right)^{1/2} \leq y \leq \left(R^2 - x^2\right)^{1/2} \right]$

$$= \frac{2}{\pi R^2} \int_0^{\sqrt{R^2 - x^2}} 1 \cdot dy = \frac{2}{\pi R^2} \left(R^2 - x^2\right)^{1/2}$$

$$= \frac{2}{\pi R} \left[1 - \left(\frac{x}{R}\right)^2 \right]^{1/2}$$

Example 5·28. *Given:*

$$f(x, y) = e^{-(x+y)} \; I_{(0,\infty)}(x) \, . \, I_{(0,\infty)}(y) \, ,$$

find (i) $P(X > 1)$, (ii) $P(X < Y \mid X < 2Y)$, (iii) $P(1 < X + Y < 2)$

[Delhi Univ. B.Sc. (Maths Hons.), 1987]

Solution. We are given :

$$f(x, y) = e^{-(x+y)} \; ; \quad 0 \leq x < \infty, \; 0 \leq y < \infty \qquad \text{...(1)}$$
$$= \left(e^{-x}\right)\left(e^{-y}\right)$$
$$= f_X(x) \, . \, f_Y(y) \; ; \quad 0 \leq x < \infty, \; 0 \leq y < \infty$$

$\Rightarrow$ X and Y are independent and

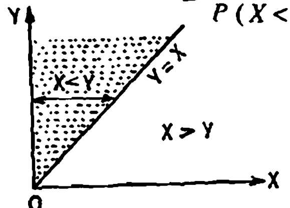
$$f_X(x) = e^{-x} \; ; \; x \geq 0 \quad \text{and} \quad f_Y(y) = e^{-y} \; ; \; y \geq 0 \qquad \text{...(2)}$$

(i) $$P(X>1)=\int_1^{\infty} f_X(x)\,dx=\int_1^{\infty} e^{-x}\,dx$$

$$=\left|\frac{e^{-x}}{-1}\right|_1^{\infty}=\frac{1}{e}$$

(ii) $$P(X<Y \mid X<2Y)=\frac{P(X<Y \cap X<2Y)}{P(X<2Y)}$$

$$=\frac{P(X<Y)}{P(X<2Y)} \qquad \dots (3)$$



$$P(X<Y)=\int_0^{\infty}\left[\int_0^{y} f(x,y)\,dx\right]dy$$

$$=\int_0^{\infty}\left[e^{-y}\left|\frac{e^{-x}}{-1}\right|_0^{y}\right]dy=-\int_0^{\infty} e^{-y}\left(e^{-y}-1\right)dy$$

$$=-\left|\frac{e^{-2y}}{-2}+e^{-y}\right|_0^{\infty}=1-\frac{1}{2}=\frac{1}{2}$$

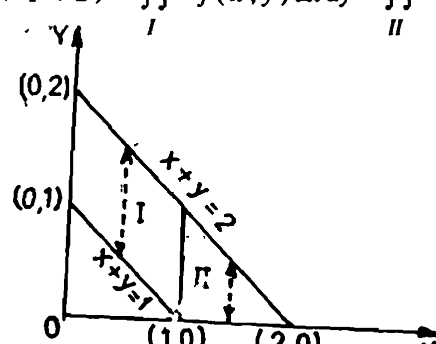
$$P(X<2Y)=\int_0^{\infty}\left[\int_0^{2y} f(x,y)\,dx\right]dy=-\int_0^{\infty} e^{-y}(e^{-2y}-1)\,dy$$

$$=-\left|\frac{e^{-3y}}{-3}+e^{-y}\right|_0^{\infty}=1-\frac{1}{3}=\frac{2}{3}$$

Substituting in (3),

$$P(X<Y \mid X<2Y)=\frac{1/2}{2/3}=\frac{3}{4}$$

(iii) $$P(1<X+Y<2)=\iint_{I} f(x,y)\,dx\,dy=\iint_{II} f(x,y)\,dx\,dy$$



$$= \int_0^1 \left(\int_{1-x}^{2-x} f(x,y)\,dy \right) dx + \int_1^2 \left(\int_0^{2-x} f(x,y)\,dy \right) dx$$

$$= \int_0^1 \left(e^{-x} \int_{1-x}^{2-x} e^{-y}\,dy \right) dx + \int_1^2 \left(e^{-x} \int_0^{2-x} e^{-y}\,dy \right) dx$$

$$= \int_0^1 \frac{e^{-x}}{-1}\left(e^{x-2} - e^{x-1}\right) dx + \int_1^2 \frac{e^{-x}}{-1}\left(e^{x-2} - 1\right) dx$$

$$= -\left(e^{-2} - e^{-1}\right) \int_0^1 1 \cdot dx - \int_1^2 \left(e^{-2} - e^{-x}\right) dx$$

$$= -\left(e^{-2} - e^{-1}\right) \Big| x \Big|_0^1 - \Big| e^{-2} \cdot x + e^{-x} \Big|_1^2$$

$$= 2/e - 3/e^2$$

Example 5·29. *(i) Let $F(x, y)$ be the d.f. of X and Y. Show that*
$$P(a < X \le b,\ c < Y \le d) = F(b,d) - F(b,c) - F(a,d) + F(a,c)$$
where a, b, c, d are real constants $a < b$; $c < d$.

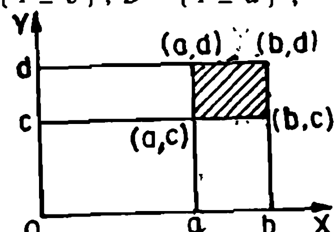
Deduce that if: $F(x,y) = 1$, for $x + 2y \ge 1$
$F(x,y) = 0$, for $x + 2y < 1$,

then $F(x, y)$ cannot be joint distribution function of variables X and Y.

(ii) Show that, with usual notation : for all x, y,
$$F_X(x) + F_Y(y) - 1 \le F_{XY}(x,y) \le \sqrt{F_X(x)\,F_Y(y)}$$
[Delhi Univ. B.Sc. (Maths Hons.), 1985]

Solution. *(i)* Let us define the events :

$A : \{X \le a\}$; $B : \{X \le b\}$; $C = \{Y \le c\}$; $D = \{Y \le d\}$;



for $a < b$; $c < d$.

$P(a < X \le b \cap c < Y \le d)$
$= P[(B - A) \cap (D - C)]$
$= P[B \cap (D - C) - A \cap (D - C)]$ ...(*)

(By distributive property of sets)

We know that if $E \subset F \Rightarrow E \cap F = E$, then

$P(F - E) = P(\bar{E} \cap F) = P(F) - P(E \cap F) = P(F) - P(E)$ ... (**)

Obviously $A \subset B \Rightarrow [A \cap (D - C)] \subset [B \cap (D - C)]$

Hence using (**), we get from (*)

$P(a < X \le b \cap c < Y \le d) = P[B \cap (D - C)] - P[A \cap (D - C)]$
$= P[(B \cap D) - (B \cap C)] - P[(A \cap D) - (A \cap C)]$

$$= P(B \cap D) - P(B \cap C) - P(A \cap D) + P(A \cap C) \quad \ldots (***)$$

[On using (**), since $C \subset D \Rightarrow (B \cap C) \subset (B \cap D)$ and $(A \cap C) \subset (A \cap D)$]

We have :

$P(B \cap D) = P[X \le b \cap Y \le d] = F(b, d)$.

Similarly

$P(B \cap C) = F(b, c)$; $P(A \cap D) = F(a, d)$ and $P(A \cap C) = F(a, c)$

Substituting in (***), we get :

$$P(a < X \le b \cap c < Y \le d) = F(b, d) - F(b, c) - F(a, d) + F(a, c) \ldots (1)$$

We are given

$$\left.\begin{aligned} F(x, y) &= 1, \quad \text{for } x + 2y \ge 1 \\ &= 0, \quad \text{for } x + 2y < 1 \end{aligned}\right\} \quad \ldots (2)$$

In (1) let us take : $a = 0$, $b = 1/2$, ; $c = 1/4$, $d = 3/4$ s.t. $a < b$ and $c < d$. Then using (2) we get :

$F(b, d) = 1$; $F(b, c) = 1$; $F(a, d) = 1$; $F(a, c) = 0$.

Substituting in (1) we get :

$P(a < X \le b \cap c < Y \le d) = 1 - 1 - 1 + 0 = -1$;

which is not possible since $P(\,.\,) \ge 0$.

Hence $F(x, y)$ defined in (2) cannot be the distribution function of variates X and Y.

(ii) Let us define the events : $A = \{X \le x\}$; $B = \{Y \le y\}$

$$\left.\begin{aligned} &\text{Then } P(A) = P(X \le x) = F_X(x) ; \ P(B) = P(Y \le y) = F_Y(y) \\ &\text{and } P(A \cap B) = P(X \le x \cap Y \le y) = F_{XY}(x, y) \end{aligned}\right\} \ldots (3)$$

$(A \cap B) \subset A \Rightarrow P(A \cap B) \le P(A) \Rightarrow F_{XY}(x, y) \le F_X(x)$

$(A \cap B) \subset B \Rightarrow P(A \cap B) \le P(B) \Rightarrow F_{XY}(x, y) \le F_Y(y)$

Multiplying these inequalities we get :

$$F^2_{X,Y}(x, y) \le F_X(x) F_Y(y) \Rightarrow F_{XY}(x, y) \le \sqrt{F_X(x) F_Y(y)} \quad \ldots (4)$$

Also $P(A \cup B) \le 1 \Rightarrow P(A) + P(B) - P(A \cap B) \le 1$

$\Rightarrow P(A) + P(B) - 1 \le P(A \cap B)$

$$\Rightarrow F_X(x) + F_Y(y) - 1 \le F_{XY}(x, y) \quad \ldots (5)$$

From (4) and (5) we get :

$F_X(x) + F_Y(y) - 1 \le F_{XY}(x, y) \le \sqrt{F_X(x) F_Y(y)}$, as required.

Example 5·30. *If X and Y are two random variables having joint density function*

$$f(x, y) = \frac{1}{8}(6 - x - y) \ ; \ 0 < x < 2, \ 2 < y < 4$$
$$= 0, \ \textit{otherwise}$$

Find (i) $P(X < 1 \cap Y < 3)$, *(ii)* $P(X + Y < 3)$ *and (iii)* $P(X < 1 \mid Y < 3)$

(Madras Univ. B.Sc., Nov. 1986)

Solution. We have

(i) $$P(X<1 \cap Y<3) = \int_{-\infty}^{1} \int_{-\infty}^{3} f(x,y)\, dx\, dy$$

$$= \int_0^1 \int_2^3 \frac{1}{8}(6-x-y)\, dx\, dy = \frac{3}{8}$$

(ii) The probability that $X+Y$ will be less than 3 is

$$P(X+Y<3) = \int_0^1 \int_2^{3-x} \frac{1}{8}(6-x-y)\, dx\, dy = \frac{5}{24}$$

(iii) The probability that $X<1$ when it is known that $Y<3$ is

$$P(X<1 \mid Y<3) = \frac{P(X<1 \cap Y<3)}{P(Y<3)} = \frac{3/8}{5/8} = \frac{3}{5}$$

$$\left[P(Y<3) = \int_0^2 \int_2^3 \frac{1}{8}(6-x-y)\, dx\, dy = \frac{5}{8}\right]$$

Example 5·31. *If the joint distribution function of X and Y is given by :*

$$F(x,y) = 1-e^{-x}-e^{-y}+e^{-(x+y)}; \quad x>0,\ y>0$$
$$= 0; \qquad \textit{elsewhere}$$

(a) Find the marginal densities of X and Y.

(b) Are X and Y independent ?

(c) Find $P(X \le 1 \cap Y \le 1)$ *and* $P(X+Y \le 1)$. **(I.C.S., 1989)**

Solution. *(a) & (b)* The joint p.d.f. of the r.v.'s (X, Y) is given by:

$$f_{XY}(x,y) = \frac{\partial^2 F(x,y)}{\partial x\, \partial y} = \frac{\partial}{\partial x}\left[e^{-y} - e^{-(x+y)}\right]$$
$$= e^{-(x+y)}; \quad x \ge 0,\ y \ge 0$$
$$= 0; \quad \text{otherwise} \qquad \text{... (i)}$$

We have

$$f_{XY}(x,y) = e^{-x} \cdot e^{-y} = f_X(x) f_Y(y) \qquad \text{... (ii)}$$

where $f_X(x) = e^{-x};\ x \ge 0; \quad f_Y(y) = e^{-y};\ y \ge 0$ ... *(iii)*

(ii) $\Rightarrow$ X and Y are independent,

and *(iii)* gives the marginal p.d.f.'s of X and Y.

(c) $$P(X \le 1 \cap Y \le 1) = \int_0^1 \int_0^1 f(x,y)\, dx\, dy$$

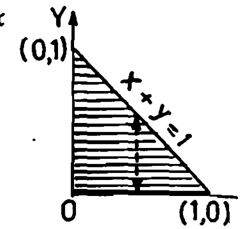
$$= \left(\int_0^1 e^{-x} dx\right)\left(\int_0^1 e^{-y} dy\right)$$

$$= \left(1-e^{-1}\right)^2$$

$$P(X+Y \le 1) = \iint_{x+y \le 1} f(x,y) = \int_0^1 \left(\int_0^{1-x} f(x,y)\, dy \right) dx$$

$$= \int_0^1 \left[e^{-x} \int_0^{1-x} e^{-y}\, dy \right] dx$$

$$= \int_0^1 e^{-x} \left(1 - e^{-(1-x)} \right) dx = 1 - 2e^{-1}$$



Example 5·32. *Joint distribution of X and Y is given by*

$$f(x,y) = 4xy\, e^{-(x^2+y^2)}; \quad x \ge 0,\ y \ge 0.$$

Test whether X and Y are independent.

For the above joint distribution, find the conditional density of X given $Y = y$. **(Calicut Univ. B.Sc., 1986)**

Solution. Joint p.d.f. of X and Y is

$$f(x,y) = 4xy\, e^{-(x^2+y^2)}; \quad x \ge 0,\ y \ge 0.$$

Marginal density of X is given by

$$f_1(x) = \int_0^\infty f(x,y)\, dy = \int_0^\infty 4xy\, e^{-(x^2+y^2)}\, dy$$

$$= 4x\, e^{-x^2} \int_0^\infty y\, e^{-y^2}\, dy$$

$$= 4x\, e^{-x^2} \cdot \int_0^\infty e^{-t} \cdot \frac{dt}{2} \qquad \left(\text{Put } y^2 = t \right)$$

$$= 2x \cdot e^{-x^2} \left| -e^{-t} \right|_0^\infty$$

$$\Rightarrow \quad f_1(x) = 2x\, e^{-x^2}; \quad x \ge 0$$

Similarly, the marginal p.d.f. of Y is given by

$$f_2(y) = \int_0^\infty f(x,y)\, dx = 2y\, e^{-y^2}; \quad y \ge 0$$

Since $f(x,y) = f_1(x) \cdot f_2(y)$, X and Y are independently distributed.

The conditional distribution of X for given Y is given by :

$$f(X = x \mid Y = y) = \frac{f(x, y)}{f_2(y)}$$
$$= 2x\, e^{-x^2}\; ;\; x \geq 0\, .$$

EXERCISE 5(e)

1. (a) Two fair dice are tossed simultaneously. Let X denote the number on the first die and Y denote the number on the second die.

(i) Write down the sample space of this experiment.

(ii) Find the following probabilities :

(1) $P(X + Y = 8)$, *(2)* $P(X + Y \geq 8)$, *(3)* $P(X = Y)$,
(4) $P(X + Y = 6 \mid Y = 4)$, *(5)* $P(X - Y = 2)$.

(Sardar Patel Univ. B.Sc., 1991)

2. *(a)* Explain the concepts *(i)* conditional probability, *(ii)* random variable, *(iii)* independence of random variables, and *(iv)* marginal and conditional probability distributions.

(b) Explain the notion of the joint distribution of two random variables. If $F(x, y)$ be the joint distribution function of X and Y, what will be the distribution functions for *the marginal distribution of X and Y ?*

What is meant by the *conditional distribution* of Y under the condition that $X = x$? Consider separately the cases where *(i)* X and Y are both discrete and *(ii)* X and Y are both continuous.

3. The joint probability distribution of a pair of random variables is given by the following table :-

$Y \backslash X$	1	2	3
1	0.1	0.1	0.2
2	0.2	0.3	0.1

Find :

(i) The marginal distributions.

(ii) The conditional distribution of X given $Y = 1$.

(iii) $P\{(X + Y) < 4\}$.

4. *(a)* What do you mean by marginal and conditional distributions ? The following table represents the joint probability distribution of the discrete random variable (X, Y)

$Y \backslash X$	1	2	3
1	1/12	1/6	0
2	0	1/9	1/5
3	1/18	1/4	2/15

(i) Evaluate marginal distribution of X.

(ii) Evaluate the conditional distribution of Y given $X = 2$.

(Aligarh Univ. B.Sc., 1992)

(b) Two discrete random variables X and Y have

$$P(X=0,\ Y=0) = \frac{2}{9}\ ;\ P(X=0,\ Y=1) = \frac{1}{9}$$

$$P(X=1,\ Y=0) = \frac{1}{9}\ ;\ P(X=1,\ Y=1) = \frac{5}{9}$$

Examine whether X and Y are independent.

(Kerala Univ. B.Sc., Oct. 1987)

5. *(a)* Let the joint p.m.f. of X_1 and X_2 be

$$p(x_1, x_2) = \frac{x_1 + x_2}{21}\ ;\ x_1 = 1, 2, 3\ ;\ x_2 = 1, 2$$
$$= 0, \text{ otherwise}$$

Show that marginal p.m.f.'s of X_1 and X_2 are

$$p_1(x_1) = \frac{2x_1 + 3}{21}\ ;\ x_1 = 1, 2, 3\ ; \qquad p_2(x_2) = \frac{6 + 3x_2}{21}\ ;\ x_2 = 1, 2$$

(b) Let

$$f(x_1, x_2) = C(x_1 x_2 + e^{x_1})\ ;\ 0 < (x_1,\ x_2) < 1$$
$$= 0, \text{ elsewhere}$$

(i) Determine C.

(ii) Examine whether X_1 and X_2 are stochastically independent.

Ans. *(i)* $C = \dfrac{4}{4e - 3}$, *(ii)* $g(x_1) = C\left(\frac{1}{2}x_1 + e^{x_1}\right)$, $g(x_2) = C\left(\frac{1}{2}x_2 + e - 1\right)$

Since $g(x_1) \cdot g(x_2) \neq f(x_1, x_2)$, X_1 and X_2 are not stochastically independent.

6. Find k so that $f(x, y) = kxy$, $1 \leq x \leq y \leq 2$ will be a probability density function. **(Mysore Univ. B.Sc., 1986)**

Hint. $\iint f(x, y)\, dx\, dy = 1 \Rightarrow k \int_1^2 x \left(\int_x^2 y\, dy \right) dx = 1 \Rightarrow k = 8/9$

7. *(a)* If
$$f(x, y) = e^{-(x+y)}\ ;\ x \geq 0,\ y \geq 0$$
$$= 0, \textit{ elsewhere}$$

is the joint probability density function of random variables X and Y, find *(i)* $P(X < 1)$, *(ii)* $P(X > Y)$, and *(iii)* $P(X + Y < 1)$.

Ans. *(i)* $1 - \dfrac{1}{e}$, *(ii)* $\dfrac{1}{2}$ and *(iii)* $1 - \dfrac{2}{e}$

(b) The joint frequency function of (X, Y) is given to be

$$f(x,y) = A e^{-x-y} \; ; \quad 0 \le x \le y, \; 0 \le y < +\infty$$
$$= 0 \quad ; \quad \text{otherwise}$$

(i) Determine A.

(ii) Find the marginal density function of X.

(iii) Find the marginal density function of Y.

(iv) Examine if X and Y are independent.

(v) Find the conditional density function of Y given $X = 2$.

[Madras Univ. B.Sc. (Main Stat.), 1992]

(c) Suppose that the random variables X and Y have the joint p.d.f.

$$f(x,y) = \begin{cases} kx(x-y), & 0 < x < 2, \; -x < y < x \\ 0, & \textit{elsewhere}. \end{cases}$$

(i) Evaluate the constant k.

(ii) Find the marginal probability density functions of the random variables.

(South Gujarat Univ. B.Sc., 1988)

8. *(a)* Two-dimensional random variable (X, Y) have the joint density

$$f(x,y) = 8xy, \; 0 < x < y < 1$$
$$= 0, \; \text{otherwise}$$

(i) Find. $P(X < 1/2 \cap Y < 1/4)$.

(ii) Find the marginal and conditional distributions.

(iii) Are X and Y independent? Give reasons for your answer.

(South Gujarat Univ. B.Sc., 1992)

	$f_1(x) = 4x(1-x^2), 0 < x < 1$	$f_1(x \mid y) = 2x/y^2 \quad : 0 < x < y, 0 < y < 1$
Ans.	$= 0$, *otherwise*	
	$f_2(y) = 4y^3, 0 < y < 1$	$f_2(y \mid x) = 2y/(1-x^2) \; ; \; x < y < 1, 0 < x < 1$

9. *(a)* The random variables X and Y have the joint density function :

$$f(x,y) = 2, \; \text{if } x + y \le 1, \; x \ge 0 \text{ and } y \ge 0$$
$$= 0, \; \textit{otherwise}$$

Find the conditional distribution of Y, given $X = x$.

(Calcutta Univ. B.Sc. (Hons.), 1984)

(b) The random variables X and Y have the joint distribution given by the probability density function :

$$f(x,y) = \begin{cases} 6(1-x-y), & \text{for } x > 0, \; y > 0, \; x + y < 1 \\ 0, & \textit{elsewhere} \end{cases}$$

Find the marginal distributions of X and Y. Hence examine if X and Y are independent. **[Calcutta Univ. B.Sc. (Hons.), 1986)**

10. If the joint distribution function of X and Y is given by

$$F(x,y) = (1 - e^{-x})(1 - e^{-y}) \quad \text{for } x > 0, \; y > 0$$
$$= 0, \; \text{elsewhere}$$

Find $P(1 < X < 3,\ 1 < Y < 2)$. **[Delhi Univ. M.A.(Econ.), 1988]**

Hint. Reqd. Prob. $= \left(\int_1^3 e^{-x}\,dx\right)\left(\int_1^2 e^{-y}\,dy\right) = \left(1 - e^{-3}\right)\left(1 - e^{-2}\right)$

11. Let X and Y be two random variables with the joint probability density function

$$f(x,y) = \begin{cases} 8xy, & 0 < x \le y < 1 \\ 0 & , \textit{otherwise} \end{cases}$$

Obtain :

(i) the joint distribution function of X and Y.
(ii) the marginal probability density function of Y; and
(iii) $P\left(X \le \frac{1}{4} \mid \frac{1}{2} < Y \le 1\right)$.

12. Let X and Y be jointly distributed with p.d.f.

$$f(x,y) = \begin{cases} \frac{1}{4}(1+xy), & |x| < 1,\ |y| < 1 \\ 0 & , \textit{otherwise} \end{cases}$$

Show that X and Y are not independent but X^2 and Y^2 are independent.

Hint. $$f_1(x) = \int_{-1}^{1} f(x,y)\, dy = \frac{1}{2},\ -1 < x < 1;$$

$$f_2(y) = \int_{-1}^{1} f(x,y)\, dx = \frac{1}{2},\ -1 < y < 1$$

Since $f(x,y) \ne f_1(x) f_2(y)$, X and Y are not independent. However,

$$P(X^2 \le x) = P(|X| \le \sqrt{x}) = \int_{-\sqrt{x}}^{\sqrt{x}} f_1(x)\, dx = \sqrt{x}$$

$$\begin{aligned} P(X^2 \le x \cap Y^2 \le y) &= P(|X| \le \sqrt{x} \cap |Y| \le \sqrt{y}) \\ &= \int_{-\sqrt{x}}^{\sqrt{x}} \left[\int_{-\sqrt{y}}^{\sqrt{y}} f(u,v)\, dv\right] du \\ &= \sqrt{x} \cdot \sqrt{y} \\ &= P(X^2 \le x) \cdot P(Y^2 \le y) \end{aligned}$$

$\Rightarrow$ X^2 and Y^2 are independent.

13. *(a)* The joint probability density function of the two dimensional random variable (X, Y) is given by :

$$f(x,y)=\begin{cases} x^3y^3/16\ , & 0\le x\le 2\ ,\ 0\le y\le 2 \\ 0 & ,\ elsewhere \end{cases}$$

Find the marginal densities of X and Y. Also find the cumulative distribution functions for X and Y. **(Annamalai Univ. B.E., 1986)**

Ans. $f_X(x)=\frac{x^3}{4}\ ;\ 0\le x\le 2\ ;\quad f_Y(y)=\frac{y^3}{4}\ ;\ 0\le y\le 2$

$$F_X(x)=\begin{cases} 0 & ;\ x<o \\ x^4/16 & ;\ 0\le x\le 2 \\ 1 & ;\ x>2 \end{cases}\qquad F_Y(y)=\begin{cases} 0 & ;\ y<0 \\ y^4/16 & ;\ 0\le y\le 2 \\ 1 & ;\ y>2 \end{cases}$$

(b) The joint probability density function of the two dimensional random variable (X, Y) is given by :

$$f(x,y)=\begin{cases} \frac{8}{9}xy\ , & 1\le x\le y\le 2 \\ 0 & ,\ elsewhere \end{cases}$$

(i) Find the marginal density functions of X and Y,

(ii) Find the conditional density function of Y given $X=x$, and conditional density funciton of X given $Y=y$.

[Madras Univ. B.Sc. (Stat. Main), 1987]

Ans. *(i)* $f_X(x)=\int_x^2 f(x,y)\,dy=\frac{4}{9}x\left(4-x^2\right)\ ;\ 1\le x\le 2$

$$= 0 \qquad ;\ \text{otherwise}$$

$$f_Y(y)=\int_1^y f(x,y)\,dx=\frac{4}{9}y\left(y^2-1\right)\ ;\ 1\le y\le 2$$

$$f_{X|Y}(x\mid y)=\frac{2x}{y^2-1}\qquad ;\ 1\le x\le y$$

$$f_{Y|X}(y\mid x)=\frac{f(x,y)}{f_X(y)}=\frac{2y}{4-x^2}\ ;\ x\le y\le 2$$

14. The two random variables X and Y have, for $X=x$ and $Y=y$, the joint probability density function :

$$f(x,y)=\frac{1}{2x^2y}\ ,\ \text{for } 1\le x<\infty \text{ and } \frac{1}{x}<y<x$$

Derive the marginal distributions of X and Y. Further obtain the conditional distribution of Y for $X=x$ and also that of X given $Y=y$.

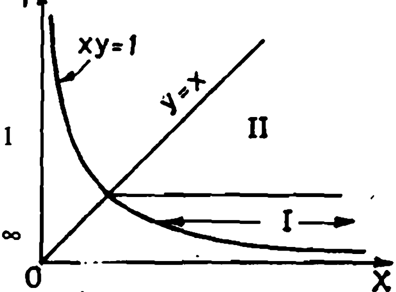
(Civil Services Main, 1986)

Hint. $f_X(x)=\int_y f(x,y)\cdot dy=\int_{1/x}^{x} f(x,y)\,dy$

$$f_Y(y) = \int_x f(x,y)\,dx$$

$$= \int_{1/y}^{\infty} f(x,y)\,dx\,; \quad 0 \le y \le 1$$

$$= \int_{y}^{\infty} f(x,y)\,dx\,; \quad 1 \le y < \infty$$



15. Show that the conditions for the function

$$f(x,y) = k \exp\left[A x^2 + 2Hxy + By^2\right], \ -\infty < (x, y) < \infty$$

to be a bivariate p.d.f. are

(i) $A \le 0$, (ii) $B \le 0$ (iii) $AB - H^2 \ge 0$.

Further show that under these conditions

$$k = \frac{1}{\pi}\left(AB - H^2\right)^{1/2}$$

Hint. $f(x,y)$ will represent the p.d.f. of a bivariate distriibution if and only if

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

$$\Rightarrow \quad k\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[Ax^2 + 2Hxy + By^2\right]dx\,dy = 1 \qquad \ldots (*)$$

We have

$$Ax^2 + 2Hxy + By^2 = A\left[x^2 + \frac{2H}{A}xy + \frac{B}{A}y^2\right]$$

$$= A\left[\left(x + \frac{H}{A}y\right)^2 + \frac{AB - H^2}{A^2}\cdot y^2\right] \qquad \ldots (**)$$

Similarly, we can write

$$Ax^2 + 2Hxy + By^2 = B\left[\left(y + \frac{H}{B}x\right)^2 + \frac{AB - H^2}{B^2}x^2\right] \qquad \ldots (***)$$

Substituting from (**) and (***) in (*) we observe that the double integral on the left hand side will converge if and only if

$$A \le 0,\ B \le 0 \ \text{and}\ AB - H^2 \ge 0,$$

as desired.

Let us take $A = -a$; $B = -b$; $H = h$ so that $AB - H^2 = ab - h^2$, where $a > 0, b > 0$.

Substituting in (*), we get

$$k \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[- \frac{ab-h^2}{a} y^2 - \frac{1}{a} (-ax + hy)^2 \right] dx\, dy = 1$$

$$\Rightarrow \qquad k \int_{-\infty}^{\infty} \left[\exp\left(- \frac{ab-h^2}{a} y^2 \right) \cdot \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{a} (ax - hy)^2 \right\} dx \right] dy$$

$$= 1 \qquad \ldots (****)$$

(By Fubini's theorem)

Now $$\int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{a} (ax - hy)^2 \right\} dx = \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{a} \right) \frac{du}{a}$$

$(ax - hy = u)$

$$= \frac{1}{a} \sqrt{\pi} \sqrt{a} = \sqrt{\frac{\pi}{a}}$$

$$\left(\because \int_{-\infty}^{\infty} e^{-c^2 u^2} du = \frac{\sqrt{\pi}}{c} \right)$$

Hence from (****), we get

$$k \sqrt{\frac{\pi}{a}} \int_{-\infty}^{\infty} \exp\left\{ - \frac{ab-h^2}{a} y^2 \right\} dy = 1$$

$$\Rightarrow \qquad k \sqrt{\frac{\pi}{a}} \cdot \sqrt{\frac{\pi a}{ab-h^2}} = 1$$

$$\Rightarrow \qquad k = \frac{1}{\pi} \sqrt{ab-h^2} = \frac{1}{\pi} \sqrt{AB-H^2}.$$

OBJECTIVE TYPE QUESTIONS

I. Which of the following statements are TRUE or FALSE.

(i) Given a continuous random variable X with probability density function $f(x)$, then $f(x)$ cannot exceed unity.

(ii) A random variable X has the following probability density function :

$$f(x) = x, \; 0 < x < 1$$
$$= 0, \text{ elsewhere}$$

(iii) The function defined as

$$f(x) = |x|, \; -1 < x < 1$$
$$= 0, \text{ elsewhere}$$

is a possible probability density function.

(iv) The following represents joint probability distribution.

Y \ X	1	2	3
−1	1/9	1/18	1/18
0	1/18	2/9	3/9
1	1/8	1/18	1/18

II. Fill in the blanks :

(i) If $p_1(x)$ and $p_2(y)$ be the marginal probability functions of two independent discrete random variablies X and Y, then their joint probability function

$$p(x, y) = \ldots$$

(ii) The function $f(x)$ defined as

$$f(x) = |x|, \; -1 < x < 1$$
$$= 0, \text{ elsewhere}$$

is a possible ......

5·6. Transformation of One-dimensional Random Variable. Let X be a random variable defined on the event space S and let $g(\cdot)$ be a function such that $Y = g(X)$ is also a r.v. defined on S. In this section we shall deal with the following problem :

"Given the probability density of a r.v. X, to determine the density of a new r.v. $Y = g(X)$."

It can be proved in general that, if $g(\cdot)$ is any continuous function, then the distribution of $Y = g(X)$ is uniquely determined by that of X. The proof of this result is rather difficult and beyond the scope of this book. Here we shall consider the following, relatively simple theorem.

Theorem 5·9. *Let X be a continuous r.v. with p.d.f. $f_X(x)$. Let $y = g(x)$ be strictly monotonic (increasing or decreasing) function of x. Assume that $g(x)$ is differentiable (and hence continuous) for all x. Then the p.d.f. of the r.v. Y is given by*

$$h_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|,$$

where x is expressed in terms of y.

Proof. Case (i). $y = g(x)$ *is strictly increasing function of x (i.e., $dy/dx > 0$*. The *d.f.* of Y is given by

$$H_Y(y) = P(Y \le y) = P[g(X) \le y] = P(X \le g^{-1}(y))],$$

the inverse exists and is unique, since $g(\cdot)$ is strictly increasing.

$$\therefore \quad H_Y(y) = F_X\left[g^{-1}(y)\right], \text{ where } F \text{ is the d.f. of } X$$

$$= F_X(x) \qquad \left[\because y = g(x) \Rightarrow g^{-1}(y) = x\right]$$

Differentiating w.r.t. y, we get

$$h_Y(y) = \frac{d}{dy}[F_X(x)] = \frac{d}{dx}\left(F_X(x)\right)\frac{dx}{dy}$$

$$= f_X(x)\frac{dx}{dy} \qquad \ldots (*)$$

Case (ii). $y = g(x)$ *is strictly monotonic decreasing.*

$$H_Y(y) = P(Y \le y) = P[g(X) \le y] = P[X \ge g^{-1}(y)]$$
$$= 1 - P\left[X \le g^{-1}(y)\right] = 1 - F_X\left[g^{-1}(y)\right] = 1 - F_X(x),$$

where $x = g^{-1}(y)$, the inverse exists and is unique. Differentiating w.r.t. y, we get

$$h_Y(y) = \frac{d}{dx}\left[1 - F_X(x)\right]\frac{dx}{dy} = -f_X(x)\cdot\frac{dx}{dy}$$
$$= f_X(x)\cdot\frac{-dx}{dy} \qquad \ldots(**)$$

Note that the algebraic sign (–ive) obtained in (**) is correct, since y is a decreasing function of $x \Rightarrow x$ is a decreasing function of $y \Rightarrow dx/dy < 0$.

The results in (*) and (**) can be combined to give

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

Example 5·33. *If the cumulative distribution function of X is F (x), find the cumulative distribution function of*

(i) $Y = X + a$, *(ii)* $Y = X - b$, *(iii)* $Y = aX$,

(iv) $Y = X^3$, and *(v)* $Y = X^2$

What are the corresponding probability density functions ?

Solution. Let $G(\cdot)$ be the c.d.f. of Y. Then

(i) $G(x) = P(Y \le x) = P[X + a \le x] = P[X \le x - a] = F(x - a)$

(ii) $G(x) = P(Y \le x) = P[X - b \le x] = P[X \le x + b] = F(x + b)$

(iii) $G(x) = P[aX \le x] = P\left[X \le \dfrac{x}{a}\right],\ a > 0$

$$= F\left(\frac{x}{a}\right), \text{ if } a > 0$$

and $$G(x) = P\left[X \ge \frac{x}{a}\right] = 1 - P\left[X < \frac{x}{a}\right]$$
$$= 1 - F\left(\frac{x}{a}\right), \text{ if } a < 0$$

(iv) $G(x) = P[Y \le x] = P[X^3 \le x] = P\left[X \le x^{1/3}\right] = F\left(x^{1/3}\right)$

(v) $G(x) = P\left[X^2 \le x\right] = \left[-x^{1/2} \le X \le x^{1/2}\right]$

$$= P\left[X \le x^{1/2}\right] - P\left[X \le -x^{1/2}\right]$$

$$= 0, \qquad \text{if } x < 0$$
$$= F(\sqrt{x}) - F(-\sqrt{x} - 0), \qquad \text{if } x > 0$$

Variable	d.f.	p.d.f.
X	$F(x)$	$f(x)$
$X - a$	$F(x+a)$	$f(x+a)$
aX	$\left.\begin{array}{l} F(x/a) \quad a > 0 \\ 1 - F(x/a), a < 0 \end{array}\right\}$	$\begin{array}{l} (1/a)\, f(x/a), \ a > 0 \\ (-1/a)\, f(x/a), \ a < 0 \end{array}$
X^2	$\left.\begin{array}{l} F(\sqrt{x}) - F(-\sqrt{x} - 0) \\ \text{for } x > 0 \\ 0, \textit{ otherwise} \end{array}\right\}$	$\left.\begin{array}{l} \frac{1}{2(\sqrt{x})}[f\sqrt{x} + f(-\sqrt{x})] \\ \text{for } x > 0 \\ = 0 \text{ for } x \le 0 \end{array}\right\}$
X^3	$F(x^{1/3})$	$\frac{1}{3} f(x^{1/3}) \cdot \frac{1}{x^{2/3}}$

EXERCISE 5(f)

1. *(a)* A random variable X has $F(x)$ as its distribution function [$f(x)$ is the density function]. Find the distribution and the density functions of the random variable :

(i) $Y = a + bX$, a and b are real numbers, *(ii)* $Y = X^{-1}$, $[P(X = 0) = 0]$, *(iii)* $Y = \tan X$, and *(iv)* $Y = \cos X$.

(b) Let $f(x) = \begin{cases} 1/2 \ , & -1 < x < 1 \\ 0 \ , & \textit{elsewhere} \end{cases}$

be the p.d.f. of the r.v. X. Find the distribution function and the p.d.f. of $Y = X^2$.

[Delhi Univ. B.Sc. (Maths Hons.), 1988]

Hint. $F(x) = P(X \le x) = \int_{-1}^{x} f(x)\, dx = \frac{1}{2}(x + 1)$ ...(*)

Distribution function $G(\cdot)$ of $Y = X^2$ is given by :

$G_Y(x) = F(\sqrt{x}) - F(-\sqrt{x}) \quad ; \quad x > 0$ [c.f. Example 5·33 *(v)*]

$= \frac{1}{2}(\sqrt{x} + 1) - \frac{1}{2}(-\sqrt{x} + 1)$ [From (*)]

$= \sqrt{x} \quad ; \quad 0 < x < 1$

(As $-1 < x < 1$, $Y = X^2$ lies between 0 and 1)

p.d.f. of $Y = X^2$ is $g(x) = G'(x) = \dfrac{1}{2\sqrt{x}} \quad ; \quad 0 < x < 1$

2. Let X be a continuous random variable with p.d.f. $f(x)$. Let $Y = X^2$. Show that the random variable Y has p.d.f. given by

$$g(y) = \begin{cases} \dfrac{1}{2\sqrt{y}} [f(\sqrt{y}) + f(-\sqrt{y})], & y > 0 \\ 0, & y \le 0 \end{cases}$$

3. Find the distribution and densitiy functions for *(i)* $Y = aX + b$, $a \ne 0$, b real, *(ii)* $Y = e^X$, assuming that $F(x)$ and $f(x)$, the distribution and the density of X are known.

Ans. *(i)* $\left.\begin{array}{ll} G(y) = F[(y-b)/a], & \text{if } a > 0 \\ G(y) = 1 - F[(y-b)/a], & \text{if } a < 0 \end{array}\right\}$ $g_1(y) = \dfrac{1}{|a|} f\left(\dfrac{y-b}{a}\right)$

(ii) $\left.\begin{array}{ll} G(y) = F(\log y), & y > 0 \\ \quad = 0, & y \le 0 \end{array}\right\}$ $\begin{array}{ll} g(y) = \frac{1}{y} f(\log y), & y > 0 \\ \quad = 0, & y \le 0 \end{array}$

4. *(a)* The random variable X has an exponential distribution

$$f(x) = e^{-x}, \quad 0 < x \le \infty$$

Find the density function of the variable *(i)* $Y = 3X + 5$, *(ii)* $Y = X^3$.

(b) Suppose that X has p.d.f.,

$$\begin{aligned} f(x) &= 2x, \quad 0 < x < 1 \\ &= 0, \quad \textit{elsewhere} \end{aligned}$$

Find the p.d.f. of $Y = 3X + 1$.

Ans. $g(y) = \frac{2}{9}(y - 1),\ 1 < y < 4$

5. Let X be a random variable with p.d.f.

$$\begin{aligned} f(x) &= \tfrac{2}{9}(x + 1) \qquad -1 < x < 2 \\ &= 0, \quad \textit{elsewhere} \end{aligned}$$

Find the p.d.f. of $U = X^2$. **[Poona Univ. B.E., 1992]**

6. Let the p.d.f. of X be

$$\begin{aligned} f(x) &= \tfrac{1}{6}, \quad -3 \le x \le 3 \\ &= 0, \quad \textit{elsewhere} \end{aligned}$$

Find the p.d.f. of $Y = 2X^2 - 3$.

7. Let X be a random variable with the distribution function :

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$$

Determine the distribution function $F_Y(y)$ of the random variable $Y = \sqrt{X}$ and hence compute mean of Y. **[Calcutta Univ. B.A.(Hons.), 1986]**

5·7. Transformation of Two-dimensional Random Variable. In this section we shall consider the problem of change of variables in the two-dimensional case. Let the *r.v.'s* U and V by the transformation $u = u(x, y)$, $v = v(x, y)$, where u and v are continuously differentiable functions for which Jacobian of transformation

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} \\ \dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v} \end{vmatrix}$$

is either > 0 or < 0 throughout the (x, y) plane so that the inverse transformation is uniquiely given by $x = x(u, v)$, $y = y(u, v)$.

Theorem 5·10. *The joint p.d.f.* $g_{UV}(u, v)$ *of the transformed variables* U *and* V *is given by*

$$g_{UV}(u, v) = f_{XY}(x, y) \,.\, |J|$$

where $|J|$ *is the modulus value of the Jacobian of transformation and* $f(x, y)$ *is expressed in terms of* u *and* v.

Proof. $P(x < X \leq x + dx,\ y < Y \leq y + dy)$

$$= P(u < U \leq u + du,\ v < V \leq v + dv)$$

$$\Rightarrow \quad f_{XY}(x, y)\, dx\, dy = g_{UV}(u, v)\, du\, dv$$

$$\Rightarrow \quad g_{UV}(u, v)\, du\, dv = f_{XY}(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du\, dv$$

$$\Rightarrow \quad g_{UV}(u, v) = f_{XY}(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| = f_{XY}(x, y)\, |J|$$

Theorem 5·11. *If* X *and* Y *are independent continuous r.v.'s, then the p.d.f. of* $U = X + Y$ *is given by*

$$h(u) = \int_{-\infty}^{\infty} f_X(v) f_Y(u - v)\, dv$$

Proof. Let $f_{XY}(x, y)$ be the joint *p.d.f.* of independent continuous *r.v.'s* X and Y and let us make the transformation :

$$u = x + y,\ v = x \quad \Rightarrow \quad x = v,\ y = u - v$$

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} \\ \dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1$$

Thus the joint p.d.f. of *r.v.'s* U and V is given by

$$g_{UV}(u, v) = f_{XY}(x, y)\, |J|$$

$$= f_X(x) \cdot f_Y(y)\, |J|$$

(Since X and Y are independent)

$$= f_X(v) \cdot f_Y(u - v)$$

The marginal density of U is given by

$$h(u) = \int_{-\infty}^{\infty} g_{UV}(u, v)\, dv$$

$$= \int_{-\infty}^{\infty} f_X(v)\, f_Y(u-v)\, dv$$

Remark. The function $h(\cdot)$ is given a special name and is said to be the *convolution* of $f_X(\cdot)$ and $f_Y(\cdot)$ and we write

$$h(\cdot) = f_X(\cdot) * f_Y(\cdot)$$

Example 5·34. *Let (X, Y) be a two-dimensional non-negative continuous r.v. having the joint density :*

$$f(x, y) = \begin{cases} 4xy\, e^{-(x^2+y^2)} & ; \; x \ge 0,\; y \ge 0 \\ 0 & , \; \textit{elsewhere} \end{cases}$$

Prove that the densitiy function of $U = \sqrt{X^2 + Y^2}$ is

$$h(u) = \begin{cases} 2u^3\, e^{-u^2} & , \; 0 \le u < \infty \\ 0 & , \; \textit{elsewhere} \end{cases}$$

[Meerut Univ. M.Sc., 1986]

Solution. Let us make the transformation :

$$u = \sqrt{x^2 + y^2} \quad \text{and} \quad v = x$$

$\Rightarrow \qquad v \ge 0,\; u \ge 0$ and $u \ge v \qquad \Rightarrow \qquad u \ge 0$ and $0 \le v \le u$

The Jacobian of transformation J is given by

$$\frac{1}{J} = \frac{\partial(u, v)}{\partial(x, y)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} \\ \dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v} \end{vmatrix} = -\cdot \frac{y}{\sqrt{x^2 + y^2}}$$

The joint p.d.f. of U and V is given by

$$g(u, v) = f(x, y)\, |J|$$

$$= 4xy\, e^{-(x^2+y^2)} \left| -\frac{\sqrt{x^2+y^2}}{y} \right|$$

$$= 4x\sqrt{x^2 + y^2}\, e^{-(x^2+y^2)}$$

$$= \begin{cases} 4vu \,.\, e^{-u^2} \; ; \; u \ge 0,\; 0 \le v \le u \\ 0 \;, \quad \textit{otherwise} \end{cases}$$

Hence the density function of $U = \sqrt{X^2 + Y^2}$ is

$$h(u) = \int_0^u g(u, v)\, dv = 4u\, e^{-u^2} \int_0^u v\, dv$$

$$= \begin{cases} 2u^3 e^{-u^2}, & u \geq 0 \\ 0, & elsewhere \end{cases}$$

Example 5·35. *Let the probability density function of the random variable (X, Y) be*

$$f(x, y) = \begin{cases} \alpha^{-2} e^{-(x+y)/\alpha} & ; \quad x, y > 0,\ \alpha > 0 \\ 0 & , \quad elsewhere \end{cases}$$

Find the distribution of $\frac{1}{2}(X - Y)$. **[Nagpur Univ. B.E., 1988]**

Solution. Let us make the transformation :

$$u = \frac{1}{2}(x - y) \quad \text{and} \quad v = y$$

$$\Rightarrow \qquad x = 2u + v \quad \text{and} \quad y = v$$

The Jacobian of the transformation is :

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 2 & 1 \\ 0 & 1 \end{vmatrix} = 2$$

Thus, the joint p.d.f. of the random variables (U, V) is given by :

$$g(u, v) = \begin{cases} \dfrac{2}{\alpha^2} e^{-(2/\alpha)(u+v)}, & -\infty < u < \infty,\ v > -2u, \text{ if } u < 0 \\ & v > 0 \text{ if } u \geq 0 \text{ and } \alpha > 0 \\ 0, & elsewhere \end{cases}$$

The marginal p.d.f. of U is given by

$$g_U(u) = \begin{cases} \displaystyle\int_{-2u}^{\infty} \frac{2}{\alpha^2} \exp\{-(2/\alpha)(u+v)\}\, dv \\ \quad = \dfrac{1}{\alpha} e^{-2u/\alpha}, \quad u < 0 \\ \displaystyle\int_0^{\infty} \frac{2}{\alpha} \exp\{-(2/\alpha)(u+v)\}\, dv \\ \quad = \dfrac{1}{\alpha} e^{-2u/\alpha}, \quad u \geq 0 \end{cases}$$

Hence

$$g_U(u) = \frac{1}{\alpha} e^{-(2/\alpha)|u|} \quad ; \quad -\infty < u < \infty$$

Example 5·36. *Given the joint density function of X and Y as*

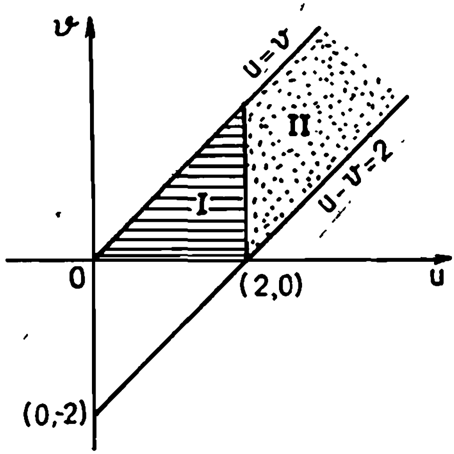
$$f(x,y)=\frac{1}{2}x\,e^{-y}\ ;\ 0<x<2,\ y>0$$
$$=0,\ \textit{elsewhere}$$

Find the distribution of $X+Y$.

Solution. Let us make the transformation :

$$u=x+y \text{ and } v=y \quad\Rightarrow\quad y=v,\ x=u-v$$

The Jacobian of transformation $J=\dfrac{\partial(x,y)}{\partial(u,v)}=1$ and the region $0<x<2$ and $y>0$ transforms to $0<u-v<2$ and $v>0$ as shown in the following figure.



The joint density function of U and V is given by

$$g(u,v)=\frac{1}{2}(u-v)\,e^{-v}\ ;\ 0<v<u,\ u>0$$

To find the density of $U=X+Y$, we split the range of U into two parts (i) $0<u\le 2$ (region I) (ii) $u>2$ (region II) (which is suggested by the diagram).

For $0<u\le 2$, (Region I) :

$$h(u)=\int_0^u g(u,v)\,dv=\frac{1}{2}\int_0^u (u-v)\,e^{-v}\,dv$$

$$=\frac{1}{2}\left|-e^{-v}(u-v)+e^{-v}\right|_{v=0}^{v=u} \qquad \text{(Integration by parts)}$$

$$=\frac{1}{2}\left(e^{-u}+u-1\right)$$

For $2 < u < \infty$, (Region II):

$$h(u) = \frac{1}{2}\int_{u-2}^{u} (u-v)\, e^{-v}\, dv$$

$$= \frac{1}{2}\left| e^{-v}(1+v-u)\right|_{v=u-2}^{v=u}$$

$$= \frac{1}{2} e^{-u}\left(1+e^{2}\right) \qquad \text{(on simplification)}$$

Hence

$$g(u) = \begin{cases} \frac{1}{2}\left(e^{-u} + u - 1\right), & 0 < u \le 2 \\ \frac{1}{2} e^{-u}\left(1+e^{2}\right), & 2 < u < \infty \\ 0, & \textit{elsewhere} \end{cases}$$

MISCELLANEOUS EXERCISE ON CHAPTER FIVE

1. 4 coins are tossed. Let X be the number of heads and Y be the number of heads minus the number of tails. Find the probability function of X, the probability function of Y and $P(-2 \le Y < 4)$.

Ans. Probability function of X is

Values of X, x	0	1	2	3	4
$p_1(x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Probability function of Y is

Values of Y, y	4	2	0	-2	-4
$p_2(y)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

$$P(-2 \le Y < 4) = \frac{4+6+4}{16} = \frac{7}{8}.$$

2. A random process gives measurements X between 0 and 1 with a probability density function

$$f(x) = 12x^3 - 21x^2 + 10x, \quad 0 \le x \le 1$$
$$= 0, \quad \textit{elsewhere}$$

(i) Find $P(X \le \frac{1}{2})$ and $P(X > \frac{1}{2})$

(ii) Find a number k such that $P(X \le k) = \frac{1}{2}$.

Ans. *(i)* 9/16, 7/16, *(ii)* $k = 0{\cdot}452$.

3. Show that for the distribution

$$dF = y_o \left[1 - \frac{|x-b|}{a} \right] dx, \quad b-a < x < b+a$$
$$= 0, \quad otherwise,$$
$$y_o = \frac{1}{a}, \quad mean = b \text{ and } variance = a^2/6$$

4. A ray of light is sent in a random direction towards the x–axis from a station Q (0, 1) on the y–axis and the ray meets the x–axis at a point P. Find the probability density function of the abscissa of P.

[Calcutta Univ. B.Sc.(Hons.), 1982]

5. Let X be a continuous variate with p.d.f.

$$f(x) = k(x - x^2) \; ; \quad a < x < b, \; k > 0$$

What are the possible values of a and b and what is k ?

[Delhi Univ. B.Sc.(Maths Hons.), 1989]

6. Pareto distribution with parameters r and A is given by the probability density function

$$f(x) = r A^r \frac{1}{x^{r+1}}, \quad \text{for } x \geq A$$
$$= 0, \; x < A, \; r > 0$$

Show that it has a finite nth moment if and only if $n < r$. Find the mean and variance of the distribution.

7. For a continuous random variable X, defined in the range $(0 \leq x < \infty)$, the probability distribution is such that

$$P(X \leq x) = 1 - e^{-\beta x^2}, \; where \; \beta > 0$$

Find the median of the distribution. Also if m, m_o and σ denote the mean, mode and standard deviation respectively of the distribution, prove that

$$2m_o^2 - m^2 = \sigma^2 \quad \text{and} \quad m_o = m\sqrt{2/\pi}$$

What is the sign of skewness of the distribution ?

8. *(a)* Two dice are rolled, $S = \{(a, b) \mid a, b = 1, 2 \ldots, 6\}$. Let X denote the sum of the two faces and Y the absolute value of their difference, *i.e.*, X is distributed over the integers 2, 3, ..., 12 and Y over 0, 1, 2, ..., 5. Assuming the dice are fair, find the probabilities that *(i)* $X = 5 \cap Y = 1$, *(ii)* $X = 7 \cap Y \geq 3$, *(iii)* $X = Y$, and *(iv)* $X + Y = 4 \cap X - Y = 2$.

Ans. *(i)* $1/8$, *(ii)* $1/9$, *(iii)* 0 and *(iv)* $1/18$.

9. The joint probability density function of the two-dimensional variable (X, Y) is of the form

$$f(x, y) = k e^{-(x+y)}, \quad 0 \leq y < x < \infty$$
$$= 0, \; elsewhere$$

(i) Determine the constant k. *(ii)* Find the conditional probability density function $f_1(x \mid y)$ and *(iii)* Compute $P(Y \geq 3)$.

[Sardar Patel Univ. B.Sc., 1986]

(iv) Find the marginal frequency function $f_1(x)$ of X.

(v) Find the marginal frequency function $f_2(y)$ of Y.

(vi) Examine if X, Y are independent.

(vii) Find the conditional frequency function of Y given $X = 2$.

Ans. (i) $k = 1$, (ii) $f_1(x \mid y) = e^{-x}$, (iii) e^{-3}.

10. Let

$$f(x,y) = \begin{cases} \binom{y}{x} p^x (1-p)^{y-x} \dfrac{e^{-\lambda} \lambda^y}{y!} \\ \qquad ;\ x = 0, 1, 2, \ldots;\ y = 0, 1, 2, \ldots;\ \textit{with } y \geq x \\ 0, \ \textit{elsewhere} \end{cases}$$

Find the marginal density function of X and the marginal density function of Y. Also determine whether the random variables X and Y are independent.

[I.S.I., 1987]

11. Consider the following function :

$$f(x \mid y) = \begin{cases} \dfrac{y^x e^{-y}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \textit{otherwise} \end{cases}$$

(i) Show that $f(x \mid y)$ is the conditional probability function of X given Y; $y \geq 0$.

(ii) If the marginal p.d.f. of Y is

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda x}, & x > 0. \\ 0 & x \leq 0, \ \lambda > 0 \end{cases}$$

what is the joint p.d.f. of X and Y ?

(iii) Obtain the marginal probability function of X.

[Delhi Univ. M.A.(Econ.), 1989]

12. The probability density function of (x_1, x_2) is given as

$$f(x_1, x_2) = \begin{cases} \theta_1 \theta_2 e^{-\theta_1 x_1 - \theta_2 x_2} & \text{if } x_1, x_2 > 0 \\ 0 & \textit{otherwise}. \end{cases}$$

Find the density function of (y_1, y_2) where

$$y_1 = \frac{2x_1}{x_2} + 1, \quad y_2 = 3x_1 + x_2 \ \text{ almost everywhere.}$$

[Punjab Univ. M.A.(Econ.), 1992]

13. *(a)* Let X_1, X_2 be a random sample of size 2 from a distribution with probability density function,

$$f(x) = e^{-x}, \ 0 < x < \infty$$

$= 0$, *elsewhere*

Show

$$Y_1 = X_1 + X_2 \text{ and } Y_2 = \frac{X_1}{X_1 + X_2}$$

are independent. [Sardar Patel Univ. B.Sc., Sept. 1986]

(b) X_1, X_2, X_3 denote random sample of size 3 drawn from the distribution:

$$f(x) = e^{-x}, \quad 0 < x < \infty$$
$$= 0, \quad elsewhere$$

Show that

$$Y_1 = \frac{X_1}{X_1 + X_2}, \quad Y_2 = \frac{X_1 + X_2}{X_1 + X_2 + X_3} \quad \text{and} \quad Y_3 = X_1 + X_2 + X_3$$

are mutually independent.

14. If the probability density function of the random varaibles X and $Y \mid X$ is given by

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & elsewhere \end{cases}$$

and $$f_{Y \mid X}(y \mid x) = \begin{cases} \dfrac{e^{-x} x^y}{y!}, & y \geq 0 \\ 0, & elsewhere \end{cases}$$

respectively, find the probability density function of the random variable Y.

[Jiwaji Univ. M.Sc., 1987]

15. (a) The random variable X and Y have a joint p.d.f. $f(x, y)$ given by

$$f(x, y) = g(x + y), \quad x > 0, \ y > 0$$
$$= 0, \quad \text{otherwise.}$$

Obtain the distribution function $H(z)$ of $Z = X + Y$ and hence show that its p.d.f. is

$$h(z) = z\, g(z), \quad z > 0$$
$$= 0 \quad z \leq 0.$$

(b) The joint density function of two random variables is given by

$f(x, y) = e^{-(x+y)}$; $x > 0$, $y > 0$. Show that the p.d.f. of

$$U = \frac{X + Y}{2} \text{ is } g(u) = 4\, u\, e^{-2u}$$

[Calicut Univ. B.Sc., 1986]

16. The time X taken by a garage to repair a car is a continuous random variable with probability density function

$$f_1(x) = \begin{cases} \frac{3}{4} x(2-x), & 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$$

If, on leaving his car, a motorist goes to keep an engagement, lasting for a time Y, where Y is a continuous random variable, independent of X, with probability function

$$f_2(y) = \begin{cases} \frac{1}{2} y ; & 0 \le y \le 2 \\ 0, & elsewhere ; \end{cases}$$

determine the probability that the car will not be ready on his return.

[Calcutta Univ. B.A.(Hons.), 1988]

17. If X and Y are two independent random variables such that

$$f(x) = e^{-x}, \ x \ge 0 \quad \text{and} \quad g(y) = 3e^{-3y}, \ y \ge 0 ;$$

find the probability distribution of $Z = X/Y$.

[Madurai Univ. B.Sc., Oct. 1987]

18. The random variables X and Y are independent and their probability density functions are, respectively given by

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{\sqrt{1-x^2}}, \ |x| < 1 \quad \text{and} \quad g(y) = y e^{-y^2/2}, \ y > 0.$$

Find the joint probability density of Z and W where $Z = XY$ and $W = X$. Deduce the probability density of Z. **[Calcutta Univ. B.Sc.(Hons.), 1985]**

CHAPTER SIX

Mathematical Expectation, Generating Functions and Law of Large Numbers

6·1. Mathematical Expectation. Let X be a random variable (r.v.) with p.d.f. (p.m.f.) $f(x)$. Then its mathematical expectation, denoted by $E(X)$ is given by :

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx, \quad (\text{ for continuous } r.v.) \qquad ...(6·1)$$

$$= \sum_x x\, f(x), \quad (\text{ for discrete } r.v.) \qquad ...(6·1a)$$

provided the righthand integral or series is absolutely convergent, *i.e.*, provided

$$\int_{-\infty}^{\infty} |x f(x)|\, dx = \int_{-\infty}^{\infty} |x|\, f(x)\, dx < \infty \qquad ...(6·2)$$

or $$\sum_x |x\, f(x)| = \sum_x |x|\, f(x) < \infty \qquad ...(6·2a)$$

Remarks. 1. Since absolute convergence implies ordinary convergence, if (6·2) or (6·2*a*) holds then the integral or series in (6·1) and (6·1*a*) also exists, *i.e.*, has a finite value and in that case we define $E(X)$ by (6·1) or (6·1*a*). It should be clearly understood that although X has an expectation only if L.H.S. in (6·2) or (6·2*a*) exists, *i.e.*, converges to a finite limit, its value is given by (6·1) or (6·1*a*).

2. $E(X)$ exists iff $E|X|$ exists.

3. The expectation of a random variable is thought of as a long-term average. [See Remark to Example (6·2*a*), **page 6·19.**].

4. Expected value and variance of an Indicator Variable. Consider the indicator variable : $X = I_A$ so that

$$X = 1 \quad \text{if} \quad A \textit{ happens}$$
$$= 0 \quad \text{if} \quad \bar{A} \textit{ happens}$$

$$\therefore \quad E(X) = 1 \quad P(X=1) + 0\,.\, P(X=0)$$
$$\Rightarrow \quad E(I_A) = 1 \quad P[I_A = 1] + 0\,.\, P[I_A = 0]$$
$$\Rightarrow \quad E(I_A) = P(A)$$

This gives us a very useful tool to find $P(A)$, rather than to evaluate $E(X)$.

Thus $$P(A) = E(I_A) \qquad ...(6·2b)$$

For illustration of this result, see Example 6·14, page 6·27.

$$E(X^2) = 1^2\,.\,P(X=1) + 0^2\,.\,P(X=0) = P(I_A=1) = P(A)$$

$$\therefore \quad Var\ X = E(X^2) - [E(X)]^2 = P(A) - [P(A)]^2$$
$$= P(A)\,[1 - P(A)]$$

$$= P(A)\, P(\bar{A}) \qquad ...(6\cdot2c)$$

Illustrations. If the r.v. X takes the values $0\,!, 1\,!, 2\,!, \ldots$ with probability law

$$P(X = x\,!) = \frac{e^{-1}}{x\,!}; \quad x = 0, 1, 2, \ldots$$

then
$$\sum_{x=0}^{\infty} x\,!\; P(X = x\,!) = e^{-1} \sum_{x=0}^{\infty} 1$$

which is a divergent series. In this case $E(X)$ does not exist.

More rigorously, let us consider a random variable X which takes the values

$$x_i = (-1)^{i+1}\,(i+1); \quad i = 1, 2, 3, \ldots$$

with the probability law

$$p_i = P(X = x_i) = \frac{1}{i\,(i+1)}; \quad i = 1, 2, 3, \ldots$$

Here
$$\sum_{i=1}^{\infty} x_i\, P(X = x_i) = \sum_{i=1}^{\infty} (-1)^{i+1}\left(\frac{1}{i}\right) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots$$

Using Leibnitz test for alternating series the series on right hand side is conditionally convergent since the terms alternate in sign, are monotonically decreasing and converge to zero. By conditional convergence we mean that although $\sum_{i=1}^{\infty} p_i x_i$ converges, $\sum_{i=1}^{\infty} |p_i\, x_i|$ does not converge. So, rigorously speaking, in the above example $E(X)$ does not exist, although $\sum_{i=1}^{\infty} p_i x_i$ is finite, *viz.*, $\log_e 2$.

As another example, let us consider the r.v. X which takes the values

$$x_k = \frac{(-1)^k \cdot 2^k}{k}; \quad k = 1, 2, 3, \ldots$$

with probabilities $p_k = 2^{-k}$.

Here also we get

$$\sum_{k=1}^{\infty} x_k\, p_k = \sum_{k=1}^{\infty} \frac{(-1)^k}{k}$$
$$= -\left[1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots\right] = -\log_e 2$$

and
$$\sum_{k=1}^{\infty} |x_k|\, p_k = \sum_{k=1}^{\infty} \frac{1}{k},$$

which is a divergent series. Hence in this case also expectation does not exist.

As an illustration of a continuous r.v. let us consider the r.v. X with p.d.f.

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2} \quad ; \quad -\infty < x < \infty$$

which is p.d.f. of Standard Cauchy distribution. [c.f.s. 8·9].

$$\int_{-\infty}^{\infty} |x| f(x)\, dx = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{|x|}{1+x^2}\, dx = \frac{2}{\pi} \int_{0}^{\infty} \frac{x}{1+x^2}\, dx$$

($\because$ *Integrand is an even function of x*)

$$= \frac{1}{\pi} \left| \log\left(1+x^2\right) \right|_0^{\infty} \to \infty$$

Since this integral does not converge to a finite limit, $E(X)$ does not exist.

6·2. Expectation of a Function of a Random Variable. Consider a r.v. X with p.d.f. (p.m.f.) $f(x)$ and distribution function $F(x)$. If $g(\cdot)$ is a function such that $g(X)$ is a r.v. and $E[g(X)]$ exists (*i.e.*, is defined), then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, dF(x) = \int_{-\infty}^{\infty} g(x) f(x)\, dx \qquad ...(6·3)$$

(*For continuous r.v.*)

$$= \sum_x g(x) f(x) \qquad ...(6·3a)$$

(*For discrete r.v.*)

By definition, the expectation of $Y = g(X)$ is

$$E[g(X)] = E(Y) = \int y \cdot dH_Y(y) = \int y\, h(y)\, dy \qquad ...(6·4)$$

or
$$E(Y) = \sum_y y\, h(y) \qquad ...(6·4a)$$

where $H_Y(y)$ is the distribution function of Y and $h(y)$ is p.d.f. of Y.

[The proof of equivalence of (6·3) and (6·4) is beyond the scope of the book.]

This result extends into higher dimensions. If X and Y have a joint p.d.f. $f(x, y)$ and $Z = h(x, y)$ is a random variable for some function h and if $E(Z)$ exists, then

$$E(Z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y)\, dx\, dy \qquad ...(6·5)$$

or
$$E(Z) = \sum_x \sum_y h(x, y)\, f(x, y) \qquad ...(6·5a)$$

Particular Cases. 1. If we take $g(X) = X^r$, r being a positive integer, in (6·3) we get :

$$E(X^r) = \int_{-\infty}^{\infty} x^r \cdot f(x)\, dx \qquad ...(6·5b)$$

which is defined as μ_r', the rth moment (about origin) of the probability distribution.

Thus μ_r' (*about origin*) $= E(X^r)$. In particular

μ_1' (*about origin*) $= E(X)$ and μ_2' (*about origin*) $= E(X^2)$.

Hence $\quad Mean = \bar{x} = \mu_1' \text{ (about origin) } = E(X)$ ...(6·6)

and $\quad \mu_2 = \mu_2' - \mu_1'^2 = E(X^2) - \{E(X)\}^2$ ...(6·6a)

2. If $g(X) = [X - E(X)]^r = (X - \bar{x})^r$, then from (6·3) we get :

$$E[X - E(X)]^r = \int_{-\infty}^{\infty} [x - E(X)]^r f(x)\, dx = \int_{-\infty}^{\infty} (x - \bar{x})^r f(x)\, dx \qquad ...(6·7)$$

which is μ_r, the rth moment about mean.

In particular, if $r = 2$, we get

$$\mu_2 = E[X - E(X)]^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 f(x)\, dx \qquad ...(6·8)$$

Formulae (6·6a) and (6·8) give the variance of the probability distribution of a r.v. X in terms of expectation.

3. Taking $g(x) = constant = c$, say in (6·3) we get

$$E(c) = \int_{-\infty}^{\infty} c\,.f(x)\, dx = c \int_{-\infty}^{\infty} f(x)\, dx = c \qquad ...(6·9)$$

$$E(c) = c \qquad ...(6·9a)$$

Remark. The corresponding results for a discrete r.v. X can be obtained on replacing integration by summation (Σ) over the given range of the variable X in the formulae (6·5) to (6·9).

In the following sections, we shall establish some more results on Expectation in the form of theorems, for continuous r.v.'s only. The corresponding results for discrete r.v.'s can be obtained similarly on replacing integration by summation (Σ) over the given range of the variable X and are left as an exercise to the reader.

6·3. Addition Theorem of Expectation

Theorem 6·1. *If X and Y are random variables then*

$$E(X + Y) = E(X) + E(Y), \qquad ...(6·10)$$

provided all the expectations exist.

Proof. Let X and Y be continuous r.v.'s with joint p.d.f. $f_{X,Y}(x, y)$ and marginal p.d.f 's $f_X(x)$ and $f_Y(y)$ respectively. Then by definition :

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \qquad ...(6·11)$$

$$E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy \qquad ...(6·12)$$

$$E(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{XY}(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{XY}(x, y)\, dx\, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f_{XY}(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \right] dx + \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \right] dy$$

$$= \int_{-\infty}^{\infty} x\, f_X(x)\, dx + \int_{-\infty}^{\infty} y\, f_Y(y)\, dy$$

$$= E(X) + E(Y) \qquad \text{[On using (6·11) and (6·12)]}$$

The result in (6·10) can be extended to n variables as given below.

Theorem 6·1(a). *The mathematical expectation of the sum of n random variables is equal to the sum of their expectations, provided all the expectations exist.*

Symbolically, if $X_1, X_2, \ldots, X_n$ are random variables then

$$E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n) \qquad \ldots(6·13)$$

or $$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i), \qquad \ldots(6·13a)$$

if all the expectations exist.

Proof. Using (6·10), for two r.v.'s X_1 and X_2 we get :

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

$\Rightarrow$ (6·13) is true for $n = 2$. ...(*)

Let us now suppose that (6·13) is true for $n = r$ (say), so that

$$E\left(\sum_{i=1}^{r} X_i\right) = \sum_{i=1}^{r} E(X_i) \qquad \ldots(6·14)$$

$$E\left(\sum_{i=1}^{r+1} X_i\right) = E\left[\sum_{i=1}^{r} X_i + X_{r+1}\right]$$

$$= E\left(\sum_{i=1}^{r} X_i\right) + E(X_{r+1}) \qquad \text{[Using (6·10)]}$$

$$= \sum_{i=1}^{r} E(X_i) + E(X_{r+1}) \qquad \text{[Using (6·14)]}$$

$$= \sum_{i=1}^{r+1} E(X_i)$$

Hence if (6·13) is true for $n = r$, it is also true for $n = r + 1$. But we have proved in (*) above that (6·13) is true for $n = 2$. Hence it is true for $n = 2 + 1 = 3$; $n = 3 + 1 = 4$; ... and so on. Hence by the principle of mathematical Introduction (6·13) is true for all positive integral values of n.

6·4. Multiplication Theorem of Expectation

Theorem 6·2. *If X and Y are independent random variables, then*

$$E(XY) = E(X) \cdot E(Y) \qquad \ldots(6\cdot15)$$

Proof. Proceeding as in Theorem 6·1, we have :

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, y\, f_{XY}(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, y\, f_X(x)\, f_Y(y)\, dx\, dy$$

[Since X and Y are independent]

$$= \int_{-\infty}^{\infty} x\, f_X(x)\, dx \int_{-\infty}^{\infty} y\, f_Y(y)\, dy$$

$$= E(X) E(Y),$$ *[Using (6·11) and (6·12)]*

provided X and Y are independent.

Generalisation to n-variables.

Theorem 6·2(a). *The mathematical expectation of the product of a number of independent random variables is equal to the product of their expectations. Symbolically, if $X_1, X_2, \ldots, X_n$ are n independent random variables, then*

$$\left.\begin{aligned} E(X_1 X_2 \ldots X_n) &= E(X_1) E(X_2) \ldots E(X_n) \\ \text{i.e.,} \quad E\left(\prod_{i=1}^{n} X_i\right) &= \prod_{i=1}^{n} E(X_i) \end{aligned}\right\} \qquad \ldots(6\cdot16)$$

provided all the expectations exist.

Proof. Using (6·15), for two *independent* random variables X_1 and X_2, we get:

$$E(X_1 X_2) = E(X_1) E(X_2).$$

$\Rightarrow$ (6·16) is true for $n = 2$. ...(*)

Let us now suppose that (6·16) is true for $n = r$, (say) so that :

$$E\left(\prod_{i=1}^{r} X_i\right) = \prod_{i=1}^{r} E(X_i) \qquad \ldots(6\cdot17)$$

$$E\left(\prod_{i=1}^{r+1} X_i\right) = E\left(\prod_{i=1}^{r} X_i \times X_{r+1}\right)$$

$$= E\left(\prod_{i=1}^{r} X_i\right) E\left(X_{r+1}\right) \qquad \text{[Using (6·15)]}$$

$$= \prod_{i=1}^{r} \left(E X_i\right) E\left(X_{r+1}\right) \qquad \text{[Using (6·17)]}$$

$$= \prod_{i=1}^{r+1} \left(E X_i\right)$$

Hence if (6·16) is true for $n = r$, it is also true for $n = r + 1$. Hence using (*), by the principle of mathematical induction we conclude that (6·16) is true for all positive integral values of n.

Theorem 6·3. *If X is a random variable and 'a' is constant, then*

(i) $E[a\Psi(X)] = a\,E[\Psi(X)]$ ...(6·18)

(ii) $E[\Psi(X) + a] = E[\Psi(X)] + a,$ ...(6·19)

where $\Psi(X)$, a function of X, is a r.v. and all the expectations exist.

Proof.

(i) $$E[a\,\Psi(X)] = \int_{-\infty}^{\infty} a\,\Psi(x)\cdot f(x)\,dx = a\int_{-\infty}^{\infty}\Psi(x) f(x)\,dx = a\,E[\Psi(X)]$$

(ii) $$E[\Psi(X) + a] = \int_{-\infty}^{\infty}[\Psi(x) + a]\,f(x)\,dx$$

$$= \int_{-\infty}^{\infty}\Psi(x)\,f(x)\,dx + a\int_{-\infty}^{\infty} f(x)\,dx$$

$$= E[\Psi(X)] + a \qquad \left(\because \int_{-\infty}^{\infty} f(x)\,dx = 1\right)$$

Cor. *(i)* If $\Psi(X) = X$, then

$E(aX) = a\,E(X)$ and $E(X + a) = E(X) + a$ ...(6·20)

(ii) If $\Psi(X) = 1$, then $E(a) = a$. ...(6·21)

Theroem 6·4. *If X is a random variable and a and b are constants, then*

$$E(aX + b) = a\,E(X) + b \qquad ...(6·22)$$

provided all the expectations exist.

Proof. By definition, we have

$$E(aX + b) = \int_{-\infty}^{\infty}(ax + b)\,f(x)\,dx$$

$$= a\int_{-\infty}^{\infty} x\,f(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx$$

$$= a\,E(X) + b$$

Cor. 1. If $b = 0$, then we get

$$E(aX) = a\cdot E(X) \qquad ...(6·22a)$$

Cor. 2. Taking $a = 1$, $b = -\bar{X} = -E(X)$, we get

$$E(X - \bar{X}) = 0$$

Remark. If we write,

$$g(X) = aX + b \qquad ...(6·23)$$

then $$g[E(X)] = a\,E(X) + b \qquad ...(6·23a)$$

Hence from (6·22) and (6·23a) we get

$$E[g(X)] = g[E(X)] \qquad \text{....(6·24)}$$

Now (6·23) and (6·24) imply *that expectation of a linear function is the same linear function of the expectation.* The result, however, is not true if $g(\cdot)$ is not linear. For instance

$$E(1/X) \neq (1/EX) \quad ; \quad E(X^{\frac{1}{2}}) \neq [E(X)]^{\frac{1}{2}},$$
$$E[\log(X) \neq \log[E(X)] \quad ; \quad E(X^2) \neq [E(X)]^2,$$

since all the functions stated above are non-linear. As an illustration, let us consider a random varible X which assumes only two values $+1$ and -1, each with equal probability $\frac{1}{2}$. Then

$$E(X) = 1 \times \tfrac{1}{2} + (-1) \times \tfrac{1}{2} = 0.$$

and $$E(X^2) = 1^2 \times \tfrac{1}{2} + (-1)^2 \times \tfrac{1}{2} = 1.$$

Thus $$E(X^2) \neq [E(X)]^2$$

For a non-linear function $g(X)$, it is difficult to obtain expressions for $E[g(X)]$ in terms of $g[E(X)]$, say, for $E[\log(X)]$ or $E(X^2)$ in terms of $\log[E(X)]$ or $[E(X)]^2$. However, some results in the form of inequalities between $E[g(X)]$ and $g[E(X)]$ are available, as discussed in Theorem 6·12 (Jenson's Inequality) page 6·15.

6·5. Expectation of a Linear Combination of Random Variables

Let $X_1, X_2, \ldots, X_n$ be any n random variables and if $a_1, a_2, \ldots, a_n$ are any n constants, then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i) \qquad \text{...(6·25)}$$

provided all the expectations exist.

Proof. The result is obvious from (6·13) and (6·20).

Theorem 6·5 (a). *If $X \geq 0$ then $E(X) \geq 0$.*

Proof. If X is a continuous random variable s.t. $X \geq 0$ then

$$E(X) = \int_{-\infty}^{\infty} x \cdot p(x)\,dx = \int_{0}^{\infty} x \cdot p(x)\,dx > 0,$$

$$[\because \text{ If } X \geq 0,\ p(x) = 0 \text{ for } x < 0]$$

provided the expectation exists.

Theorem 6·5 (b). *Let X and Y be two random variables such that $Y \leq X$ then*

$$E(Y) \leq E(X),$$

provided the expectations exist.

Proof. Since $Y \leq X$, we have the r.v.

$$Y - X \leq 0 \quad \Rightarrow \quad X - Y \geq 0$$

Hence $$E(X - Y) \geq 0 \quad \Rightarrow \quad E(X) - E(Y) \geq 0$$

$$\Rightarrow \quad E(X) \geq E(Y) \quad \Rightarrow \quad E(Y) \leq E(X),$$

as desired.

Theorem 6·6. $|E(X)| \le E|X|$, ...(6·26)

provided the expectations exist.

Proof. Since $X \le |X|$, we have by Theorem 6·5(b)

$$E(X) \le E|X| \qquad \ldots(*)$$

Again since $-X \le |X|$, we have by Theorem 6·5(b)

$$E(-X) \le E|X|$$

$$\Rightarrow \quad -E(X) \le E|X| \qquad \ldots(**)$$

From (*) and (**), we get the desired result $|E(X)| \le E|X|$.

Theorem 6·7. *If μ_r' exists, then μ_s' exists for all $1 \le s \le r$. Mathematically, if $E(X^r)$ exists, then $E(X^s)$ exists for all $1 \le s \le r$, i.e.,*

$$E(X^r) < \infty \Rightarrow E(X^s) < \infty \quad \forall \quad 1 \le s \le r \qquad \ldots(6·27)$$

Proof. $$\int_{-\infty}^{\infty} |x|^s \, dF(x) = \int_{-1}^{1} |x|^s \, dF(x) + \int_{|x|>1} |x|^s \, dF(x)$$

If $s < r$, then $|x|^s < |x|^r$ for $|x| > 1$.

$$\therefore \int_{-\infty}^{\infty} |x|^s \, dF(x) \le \int_{-1}^{1} |x|^s \, dF(x) + \int_{|x|>1} |x|^r \, dF(x)$$

$$\le \int_{-1}^{1} dF(x) + \int_{|x|>1} |x|^r \, dF(x),$$

since for $-1 < x < 1$, $|x|^s < 1$.

$$\therefore \quad \int_{-\infty}^{\infty} |x|^s \, dF(x) \le 1 + E|X|^r < \infty$$

$\Rightarrow \quad E(x^s)$ exists $\forall \; 1 \le s \le r$ $\qquad [\because E(X^r)$ exists$]$

Remark. The above theorem states that if the moments of a specified order exist, then all the lower order moments automatically exist. However, the converse is not true, *i.e.*, we may have distributions for which all the moments of a specified order exist but no higher order moments exist. For example, for the r.v. with p.d.f.

$$p(x) = 2/x^3 \; ; \quad x \ge 1$$
$$= 0 \; ; \quad x < 1$$

we have :

$$E(X) = \int_{1}^{\infty} x \, p(x) \, dx = 2\int_{1}^{\infty} x^{-2} \, dx = \left| \left(\frac{-2}{x} \right) \right|_{1}^{\infty} = 2$$

$$E(X^2) = \int_1^\infty x^2\, p(x)\, dx = 2\int_1^\infty \frac{1}{x}\, dx = \infty$$

Thus for the above distribution, 1st order moment (mean) exists but 2nd order moment (variance) does not exist.

As another illustration, consider a r.v. X with p.d.f.

$$p(x) = \frac{(r+1)\, a^{r+1}}{(x+a)^{r+2}} \; ; \; x \geq 0 \; ; \quad a > 0$$

$$\mu_r' = E(X^r) = (r+1)\, a^{r+1} \int_0^\infty \frac{x^r}{(x+a)^{r+2}}\, dx$$

Put $x = ay$ and using Beta integral :

$$\int_0^\infty \frac{x^{m-1}}{(1+x)^{m+n}} = \beta(m, n),$$

we shall get on simplification :

$$\mu_r' = (r+1)\, a^r \cdot \beta(r+1, 1) = a^r$$

However,

$$\mu_{r+1}' = E\left(X^{r+1}\right) = (r+1)\, a^{r+1} \int_0^\infty \frac{x^{r+1}}{(x+a)^{r+2}}\, dx \to \infty,$$

as the intergal is not convergent. Hence in this case only the moments up to rth order exist and higher order moments do not exist.

Theorem 6·8. *If X is a random variable, then*

$$V(aX+b) = a^2 V(X), \qquad \text{...(6·28)}$$

where a and b are constants.

Proof. Let $Y = aX + b$

Then $E(Y) = aE(X) + b$

$\therefore \quad Y - E(Y) = a\{X - E(X)\}$

Squaring and taking expectation of both sides, we get

$$E\{Y - E(Y)\}^2 = a^2 E\{X - E(X)\}^2$$

$$\Rightarrow \quad V(Y) = a^2 V(X) \quad \Rightarrow \quad V(aX+b) = a^2 V(X),$$

where $V(X)$ is written for variance of X.

Cor. *(i)* If $b = 0$, then $V(aX) = a^2 V(X)$ ...(6·28*a*)

$\Rightarrow$ Variance is not independent of change of scale.

(ii) If $a = 0$, then $V(b) = 0$ ...(6·28*b*)

$\Rightarrow$ Variance of a constant in zero.

(iii) If $a = 1$, then $V(X + b) = V(X)$ ...(6·28*c*)

$\Rightarrow$ Variance is independent of change of origin.

6·6. Covariance: *If X and Y are two random variables, then covariance between them is defined as*

$$Cov\ (X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] \quad ...(6·29)$$
$$= E[XY - XE(Y) - YE(X) + E(X)E(Y)]$$
$$= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y)$$
$$= E(XY) - E(X)E(Y) \quad ...(6·29a)$$

If X and Y are independent then $E(XY) = E(X)E(Y)$ and hence in this case

$$Cov\ (X, Y) = E(X)E(Y) - E(X)E(Y) = 0 \quad ...(6·29b)$$

Remarks. 1. $Cov\ (aX, bY) = E[\{aX - E(aX)\}\{bY - E(bY)\}]$

$$= E[a\{X - E(X)\}b\{Y - E(Y)\}]$$
$$= ab\,E[\{X - E(X)\}\{Y - E(Y)\}]$$
$$= ab\ Cov\ (X, Y) \quad ...(6·30)$$

2. $Cov\ (X + a,\ Y + b) = Cov\ (X, Y)$ ...(6·30a)

3. $Cov\left(\dfrac{X - \overline{X}}{\sigma_X}, \dfrac{Y - \overline{Y}}{\sigma_Y}\right) = \dfrac{1}{\sigma_X\ \sigma_Y}\ Cov\ (X, Y)$ ...(6·30b)

4. Similarly, we shall get :

$$Cov\ (aX + b,\ cY + d) = ac\ Cov\ (X, Y) \quad ...(6·30c)$$
$$Cov\ (X + Y, Z) = Cov\ (X, Z) + Cov\ (Y, Z) \quad ...(6·30d)$$
$$Cov\ (aX + bY,\ cX + dY) = ac\sigma_X^{\ 2} + bd\sigma_Y^{\ 2} + (ad + bc)\ Cov\ (X, Y) \quad ...(6·30e)$$

If X and Y are independent, $Cov\ (X, Y) = 0$; [c.f. (6·29b)]

However, the converse is not true.

(For details see Theorem 10·2)

6·6·1. Correlation Coefficient. The correlation coefficient (ρ_{XY}), between the variables X and Y is defined as :

$$\rho_{XY} = Correlation\ Coefficient\ (X, Y) = \frac{Cov\ (X, Y)}{\sigma_X\ \sigma_Y} \quad ...(6·30f)$$

For detailed discussion on correlation coefficient, see Chapter 10.

6·7. Variance of a Linear Combination of Random Variables

Theorem 6·9. *Let $X_1, X_2, ..., X_n$ be n random variables then*

$$V\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^{\ 2}\ V(X_i) + 2\sum_{i=1}^{n}\sum_{\substack{j=1 \\ i<j}}^{n} a_i a_j\ Cov\ (X_i, X_j) \quad ...(6·31)$$

Proof. Let $U = a_1X_1 + a_2X_2 + \ldots + a_nX_n$

$\therefore \quad E(U) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$

$\therefore \quad U - E(U) = a_1[X_1 - E(X_1)] + a_2[X_2 - E(X_2)] + \ldots + a_n[X_n - E(X_n)]$

Squaring and taking expectation of both sides, we get

$$E[U - E(U)]^2 = a_1^2 E[X_1 - E(X_1)]^2 + a_2^2 E[X_2 - E(X_2)]^2 + \dots + a_n^2 E[X_n - E(X_n)]^2 + 2 \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i<j}}^{n} a_i a_j E[\{X_i - E(X_i)\}\{X_j - E(X_j)\}]$$

$$\Rightarrow \quad V(U) = a_1^2 V(X_1) + a_2^2 V(X_2) + \dots + a_n^2 V(X_n) + 2 \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i<j}}^{n} a_i a_j \, Cov(X_i, X_j)$$

$$\Rightarrow V\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 V(X_i) + 2 \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i<j}}^{n} a_i a_j \, Cov(X_i, X_j)$$

Remarks. 1. If $a_i = 1$; $i = 1, 2, \dots, n$ then

$$V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n) + 2 \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i<j}}^{n} Cov(X_i, X_j) \quad \dots(6{\cdot}31a)$$

2. If $X_1, X_2, \dots, X_n$ are independent (pairwise) then $Cov(X_i, X_j) = 0$, $(i \neq j)$. Thus from (6·31) and (6·31a), we get

$$\left.\begin{aligned} V(a_1X_1 + a_2X_2 + \dots + a_nX_n) &= a_1^2V(X_1) + a_2^2V(X_2) + \dots + a_n^2V(X_n) \\ \text{and} \quad V(X_1 + X_2 + \dots + X_n) &= V(X_1) + V(X_2) + \dots + V(X_n) \end{aligned}\right\} \quad \dots(6{\cdot}31b)$$

3. If $a_1 = 1 = a_2$ and $a_3 = a_4 = \dots = a_n = 0$, then from (6·31), we get

$$V(X_1 + X_2) = V(X_1) + V(X_2) + 2\, Cov(X_1, X_2)$$

Again if $a_1 = 1$, $a_2 = -1$ and $a_3 = a_4 = \dots = a_n = 0$, then

$$V(X_1 - X_2) = V(X_1) + V(X_2) - 2\, Cov(X_1, X_2)$$

Thus we have

$$V(X_1 \pm X_2) = V(X_1) + V(X_2) \pm 2\, Cov(X_1, X_2) \quad \dots(6{\cdot}31c)$$

If X_1 and X_2 are independent, then $Cov(X_1, X_2) = 0$ and we get

$$V(X_1 \pm X_2) = V(X_1) + V(X_2) \quad \dots(6{\cdot}31d)$$

Theorem 6·10. *If X and Y are independent random variables then*

$$E[h(X) \cdot k(Y)] = E[h(X)]\, E[k(Y)] \quad \dots(6{\cdot}32)$$

where $h(\cdot)$ is a function of X alone and $k(\cdot)$ is a function of Y alone, provided expectations on both sides exist.

Proof. Let $f_X(x)$ and $g_Y(y)$ be the marginal p.d.f.'s of X and Y respectively. Since X and Y are independent, their joint p.d.f. $f_{XY}(x, y)$ is given by

$$f_{XY}(x, y) = f_X(x)\, f_Y(y) \quad \dots(*)$$

By definition, for continuous r.v.'s

$$E[h(X) \cdot k(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)\, k(y)\, f(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)\, k(y)\, f(x)\, g(y)\, dx\, dy$$

[From (*)]

Since $E[h(X)\ k(Y)]$ exists, the integral on the right hand side is absolutely convergent and hence by Fubini's theorem for integrable functions we can change the order of integration to get

$$E[h(X)\, k(Y)] = \left[\int_{-\infty}^{\infty} h(x)\, f(x)\, dx\right]\left[\int_{-\infty}^{\infty} k(y)\, g(y)\, dy\right]$$

$$= E[h(X)] \cdot E[k(Y)],$$

as desired.

Remark. The result can be proved for discrete random variables X and Y on replacing integration by summation over the given range of X and Y.

Theorem 6·11. Cauchy-Schwartz Inequality. *If X and Y are random variables taking real values, then*

$$[E(XY)]^2 \le E(X^2) \cdot E(Y^2) \qquad \text{...(6.33)}$$

Proof. Let us consider a real valued function of the real variable t, defined by

$$Z(t) = E(X + tY)^2$$

which is always non-negative, since $(X + tY)^2 \ge 0$, for all real X, Y and t.

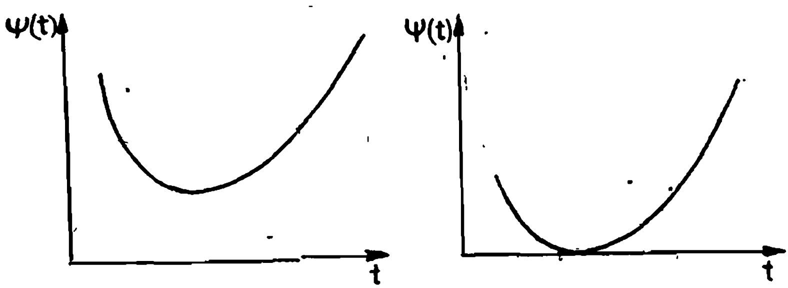
Thus $\quad Z(t) = E(X + tY)^2 \ge 0 \quad \forall \quad t.$

$\Rightarrow \quad Z(t) = E[X^2 + 2t\, XY + t^2 Y^2]$

$\qquad = E(X^2) + 2t \cdot E(XY) + t^2 E(Y^2) \ge 0$, for all t. $\quad (*)$

Obviously, $Z(t)$ is a quadratic expression in 't'.

We know that the quadratic expression of the form :



$\Psi(t) = At^2 + Bt + C \geq 0$ for all t, implies that the graph of the function $\Psi(t)$ either touches the t-axis at only one point or not at all, as exhibited in the diagrams.

This is equivalent to saying that the discriminant of the function $\Psi(t)$, *viz.*, $B^2 - 4AC \leq 0$, since the condition $B^2 - 4AC > 0$ implies that the function $\Psi(t)$ has two distinct real roots which is a contradiction to the fact that $\Psi(t)$ meets the t-axis either at only one point or not at all. Using this result, we get from (*),

$$4 \cdot [E(XY)]^2 - 4 \cdot E(X^2) \cdot E(Y^2) \leq 0$$

$$\Rightarrow \quad [E(XY)]^2 \leq E(X^2) \cdot E(Y^2)$$

Remarks. 1. The sign of equality holds in (6·33) or (*) iff

$$E(X + tY)^2 = 0 \quad \forall \ t \quad \Rightarrow \quad P[(X + tY)^2 = 0] = 1,$$

$$\Rightarrow P[X + tY = 0] = 1 \qquad \Rightarrow \quad P\left[Y = -\frac{X}{t}\right] = 1$$

$$\Rightarrow P[Y = AX] = 1 \qquad ; \quad (A = -1/t) \qquad \ldots(6·33b)$$

2. If the r.v. X takes the real values $x_1, x_2, \ldots, x_n$ and r.v. Y takes the real values $y_1, y_2, \ldots, y_n$ then Cauchy-Schwartz inequality implies :

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right)$$

$$\Rightarrow \quad \left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right) \cdot \left(\sum_{i=1}^{n} y_i^2\right),$$

the sign of equality holding if and only if :

$$\frac{x_i}{y_i} = \textit{constant} = k, \ (\textit{say}) \quad \text{for } \textit{all } i = 1, 2, \ldots, n$$

i.e. iff $\quad \dfrac{x_1}{y_1} = \dfrac{x_2}{y_2} = \cdots = \dfrac{x_n}{y_n} = k$, (*say*).

3. Replacing X by $|X - E(X)| = |X - \mu_x|$ and taking $Y = 1$ in (6·33) we get

$$[E\,|X - \mu_x|\,]^2 = E\,|X - \mu_x|^2 \cdot E(1)$$

$$\Rightarrow \quad [\textit{ Mean Deviation about mean }]^2 \leq \textit{Variance}(X)$$

$$\Rightarrow \quad M.D. \leq S.D. \ \Rightarrow \ S.D. \geq M.D. \qquad \ldots(6·33a)$$

6·7. Jenson's Inequality

Continuous Convex Function. *(Definition).* A continuous function $g(x)$ on the interval I is convex if for every x_1 and x_2, $(x_1 + x_2)/2 \in I$, we have

$$g\left(\frac{x_1 + x_2}{2}\right) \leq \frac{1}{2}g(x_1) + \frac{1}{2}g(x_2) \qquad \ldots(6·34)$$

Remarks. 1. If $x_1, x_2 \in I$, then $(x_1 + x_2)/2 \in I$.

2. Sometimes (6·34) is replaced by the stronger condition :

For $x_1, x_2 \in I$,

$$g[\lambda x_1 + (1 - \lambda) x_2] \leq \lambda g(x_1) + (1 - \lambda) g(x_2) \ ; \ 0 \leq \lambda \leq 1 \qquad \ldots(6·35)$$

(6·34) and (6·35) agree at $\lambda = \frac{1}{2}$.

3. If we do not assume the continuity of $g(x)$, then (6·35) is required to define convexity. There are certain *'non-measurable'* non-constant functions $g(\cdot)$ satisfying (6·34) but not (6·35). If $g(\cdot)$ is measurable, then (6·34) and (6·35) are equivalent.

4. A function satisfying (6·35) is continuous except possibly at the end points of the interval I (if it has end points).

5. If g is twice differentiable, *i.e.,* $g''(x)$ exists for $X \in$ [interior of I], and $g''(x) \geq 0$ for such x, then g is convex on the interior points.

6. For any point x_0 interior to I, $\exists$ a straight line $y = ax + b$, which passes through $(x_0, g(x_0))$ and satisfies $g(x) \geq ax + b$, for all $x \in I$.

Theorem 6·12. (Jenson's Inequality). *If g is continuous and convex function on the interval I, and X is a random variable whose values are in I with probability 1, then*

$$E[g(X)] \geq g[E(X)], \qquad \text{...(6·36)}$$

provided the expectations exist.

Proof. First of all we shall show that $E(X) \in I$.

The various possible cases for I are :

$$I = (-\infty, \infty);\ I = (a, \infty);\ I = [a, \infty);\ I = (-\infty, b);$$
$$I = (-\infty, b],\ I = (a, b) \text{ and variations of this.}$$

If $E(X)$ exists, then $-\infty < E(X) < \infty$.

If $X \geq a$ almost surely *(a.s.), i.e.,* with probability 1, then $E(X) \geq a$.

If $X \leq b$, a.s. then $E(X) \leq b$.

Thus $E(X) \in I$. Now $E(X)$ can be either a left end or a right end point (if end points exist) of I or an interior point of I.

Suppose I has a left end point *'a'*, *i.e.,* $X \geq a$ and $E(X) = a$. Then $X - a \geq 0$ a.s. and $E(X - a) = 0$.

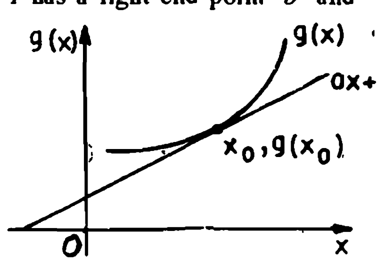
Thus $\quad P(X = a) = 1 \quad$ or $\quad P[(X - a) = 0)] = 1$.

$$\therefore \quad E[g(X)] = E[g(a)] \qquad [\because g(x) = g(a) \text{ a.s.}]$$
$$= g(a) \qquad (\because g(a) \text{ is a constant})$$
$$= g E(X)].$$

The result can be established similarly if I has a right end point *'b'* and $E(X) = b$.

Thus we are now required to establish (6·36) when $E(X) = x_0$, is an interior point of I.

Let $ax + b$ pass through the point



$(x_0, g(x_0))$ and let it be below g

[c.f. Remark 6 above].

$$\therefore \quad E[g(X)] \geq E(aX+b) = aE(X)+b$$
$$= ax_0 + b$$
$$= g(x_0) = g[E(X)]$$
$$\Rightarrow \quad E[g(X)] \geq g[(X)].$$

Continuous Concave Function. *(Definition).* A continuous function g is concave on an interval I if $(-g)$ is convex.

Corollary to Theorem 6·12. *If g is a continuous and concave function on the interval I and X is a r.v. whose values are in I with probability 1, then*

$$E[g(X)] \leq g[E(X)] \qquad \ldots(6{\cdot}37)$$

provided the expectations exist.

Remarks. 1. Equality holds in Theorem (6·12) and corollary (6·37) if and only if

$$P[g(X) = aX + b] = 1,$$

for some a and b.

2. Jenson's inequality extends to random vectors. If I is a convex set in n-dimensional Euclidean space, *i.e.*, the interval I in theorem 6·12 is transferred to convex set, g is conotinuous on I, (6·34) holds whenever X_1 and X_2 are any arbitrary vectors in I. The condition $g''(x) \geq 0$ for x interior to I implies

$$\left(\frac{\partial^2 g(x)}{\partial x_i \, \partial x_j}\right) = M(x), \quad \text{(say)},$$

is non-negative definite for all x interior to I.

SOME ILLUSTRATIONS OF JENSON'S INEQUALITY

1. If $E(X^2)$ exists, then

$$E(X^2) \geq [E(X)]^2, \qquad \ldots(6{\cdot}38)$$

since $g(X) = X^2$ is convex function of X as $g''(X) = 2 > 0$.

2. If $X > 0$ a.s. *i.e.*, X assumes only positive values and $E(X)$ and $E(1/X)$ exist then

$$E\left(\frac{1}{X}\right) \geq \frac{1}{E(X)}, \qquad \ldots(6{\cdot}38a)$$

because $g(X) = \dfrac{1}{X}$ is a convex function of X since

$$g''(X) = \frac{2}{X^3} > 0, \text{ for } X > 0.$$

3. If $X > 0$, a.s. then

$$E(X^{1/2}) \leq [E(X)]^{1/2}, \qquad \ldots(6{\cdot}38b)$$

since $g(X) = X^{1/2}$, $X > 0$ is a concave function

as $\quad g''(X) = -\frac{1}{4}X^{-3/2} < 0$, for $X > 0$.

4. If $X > 0$, a.s. then

$$E[\log(X)] \le \log[E(X)], \quad \text{...(6·38c)}$$

provided the expectations exist, because $\log X$ is a concave function of X.

5. Since $g(X) = E(e^{tX})$ is a convex function of X for all t and all X, if $E(e^{tX})$ and $E(X)$ exist then

$$E(e^{tX}) \ge e^{tE(X)} \quad \text{...(6·38d)}$$

If $\quad E(X) = 0$, then

$$M_X(t) = E(e^{tX}) \ge 1, \text{ for all } t.$$

Thus if $M_X(t)$ exists, then it has a lower bound 1, provided $E(X) = 0$. Further, this bound is attained at $t = 0$. Thus $M_X(t)$ has a minimum at $t = 0$.

6·7·1 ANOTHER USEFUL INEQUALITY. *Let f and g be monotone functions on some subset of the real line and X be a r.v. whose range is in the subset almost surely (a.s.) If the expectations exist, then*

$$E[f(X)\,g(X)] \ge E[f(X)]\;E[g(X)] \quad \text{...(6·39)}$$

or

$$E[f(X)\,g(X)] \le E[f(X)]\,.\,E[g(X)] \quad \text{...(6·39a)}$$

according as f and g are monotone in the same or in the opposite directions.

Proof. Let us consider the case when both the functions f and g are monotone in the same direction. Let x and y lie in the domain of f and g respectively.

If f and g are both monotonically increasing, then

$$y \ge x \quad \Rightarrow \quad f(y) \ge f(x) \quad \text{and} \quad g(y) \ge g(x)$$

$$\Rightarrow \quad f(y) - f(x) \ge 0 \quad \text{and} \quad g(y) - g(x) \ge 0$$

$$\Rightarrow \quad [f(y) - f(x)]\,.\,[g(y) - g(x)] \ge 0 \quad \text{...(*)}$$

If f and g are both monotonically decreasing then for $y \ge x$, we have

$$f(y) \le f(x) \quad \text{and} \quad g(y) \le g(x)$$

$$\Rightarrow \quad f(y) - f(x) \le 0 \quad \text{and} \quad g(y) - g(x) \le 0$$

$$\Rightarrow \quad [f(y) - f(x)]\,.\,[g(y) - g(x)] \ge 0 \quad \text{...(**)}$$

Hence if f and g are both monotonic in the same direction, then from (*) and (**), we get the same result, *viz.*,

$$[f(y) - f(x)]\,.\,[g(y) - g(x)] \ge 0.$$

Let us now consider independently and identically distributed (*i.i.d.*) random variables X and Y. Then from above, we get

$$E[(f(Y) - f(X))\,(g(Y) - g(X))] \ge 0.$$

$$\Rightarrow \quad E[f(Y)\,.\,g(Y)] - E[f(Y)\,.\,g(X)] - E[f(X)\,g(Y)] + E[f(X)\,.\,g(X)] \ge 0 \quad \text{...(6·40)}$$

Since X and Y are *i.i.d.* r.v.'s, we have

$$E[f(Y)\,g(Y)] = E[f(X)\,g(X)];$$

$$E[f(Y)\,g(X)] = E[f(Y)]\,E[g(X)] = E[f(X)]\,E[g(X)]$$

($\because X$ and Y are independent) $\quad$ ($\because X$ and Y are identical) and $\quad E[f(X)\,g(Y)] = E[f(X)].E[g(Y)] = E[f(X)].E[g(X)]$

Substituting in (6·40) we get

$$2\,E[f(X)\cdot g(X)] - 2\,E[f(X)]\,.\,E[g(X)] \ge 0\,,$$

$$\Rightarrow \qquad E[f(X)\cdot g(X)] \ge E[f(X)]\,.\,E[g(X)]$$

which establishes the result in (6·39).

Similarly, (6·39a) can be established, if f and g are monotonic in opposite directions, *i.e.*, if f is monotonically increasing (decreasing) and g is monotonically decreasing (increasing). The proof is left as an exercise to the reader.

SOME ILLUSTRATIONS OF INEQUALITY (6·39).

1. If X is a r.v. which takes only non-negative values, *i.e.*, if $X \ge 0$ a.s. then for $\alpha > 0$, $\beta > 0$, $f(X) = X^{\alpha}$ and $g(X) = X^{\beta}$ are monotonic in the same direction. Hence if the expectations exist,

$$E(X^{\alpha}\,.\,X^{\beta}) \ge E(X^{\alpha})\,.\,E(X^{\beta})$$

$$\Rightarrow \qquad E(X^{\alpha+\beta}) \ge E(X^{\alpha})\,E(X^{\beta})\,;\; \alpha > 0,\; \beta > 0 \qquad \ldots(6\cdot41)$$

In particular, taking $\alpha = \beta = 1$, we get

$$E(X^2) \ge [E(X)]^2\,,$$

a result already obtained in (6·38).

2. If $X \ge 0$, a.s. and $E(X^{\alpha})$ and $E(X^{-1})$ exist, then for $\alpha > 0$, we get from (6·39a)

$$E(X^{\alpha}\,.\,X^{-1}) \le E(X^{\alpha})\,.\,E(X^{-1})$$

$$\Rightarrow \qquad E(X^{\alpha})\,E\left(\frac{1}{X}\right) \ge E(X^{\alpha-1})\,;\; \alpha > 0. \qquad \ldots(6\cdot42)$$

In particular with $\alpha = 1$, we get

$$E(X)\,E\left(\frac{1}{X}\right) \ge 1,$$

a result already obtained in (6·38a)

Taking $\alpha = 2$ in (6·42), we get

$$E(X^2)\,E\left(\frac{1}{X}\right) \ge E(X)$$

$$\Rightarrow \qquad E(X^2) \ge \frac{E(X)}{E\left(\frac{1}{X}\right)} \ge \frac{1}{\left[E\left(\frac{1}{X}\right)\right]^2}\,, \qquad \ldots(6\cdot43)$$

on using (6·38a).

Taking $\alpha = 2$ and $\beta = -2$ in $f(X) = X^{\alpha}$, $g(X) = X^{\beta}$ and using (6·39a), we get

$$E(X^2)\,.\,E(X^{-2}) \ge E(X^2\,.\,X^{-2}) = 1.$$

$$\Rightarrow \qquad E(X^2) \ge \frac{1}{E(X^{-2})} \qquad \ldots(6\cdot43a)$$

which is a weaker inequality than (6·43).

3. If $M_X(t) = E(e^{tX})$ exists for all t and for some r.v. X, then

$$M_{X(u+v)} = E[e^{(u+v)X}] = E(e^{uX} \cdot e^{vX})$$
$$\geq E(e^{uX}) \cdot E(e^{vX})$$
$$= M_X(u) \cdot M_X(v)$$

$\therefore \quad M_X(u+v) \geq M_X(u) \cdot M_X(v), \quad$ for $u, v \geq 0$.

Example 6·1. *Let X be a random variable with the following probability distribution :*

x	:	-3	6	9
$P_r(X = x)$	:	$1/6$	$1/2$	$1/3$

Find $E(X)$ and $E(X^2)$ and using the laws of expectation, evaluate $E(2X+1)^2$.

(Gauhati Univ. B.Sc., 1992)

Solution. $E(X) = \Sigma x \cdot p(x)$

$$= (-3) \times \frac{1}{6} + 6 \times \frac{1}{2} + 9 \times \frac{1}{3} = \frac{11}{2}$$

$$E(X^2) = \Sigma x^2 p(x)$$
$$= 9 \times \frac{1}{6} + 36 \times \frac{1}{2} + 81 \times \frac{1}{3} = \frac{93}{2}$$

$$\therefore \quad E(2X+1)^2 = E[4X^2 + 4X + 1] = 4E(X^2) + 4E(X) + 1$$
$$= 4 \times \frac{93}{2} + 4 \times \frac{11}{2} + 1 = 209$$

Example 6·2. (a) *Find the expectation of the number on a die when thrown.*

(b) *Two unbiased dice are thrown. Find the expected values of the sum of numbers of points on them.*

Solution. *(a)* Let X be the random variable representing the number on a die when thrown. Then X can take any one of the values 1,2,3,..., 6 each with equal probability 1/6. Hence

$$E(X) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \ldots + \frac{1}{6} \times 6$$
$$= \frac{1}{6}(1 + 2 + 3 + \ldots + 6) = \frac{1}{6} \times \frac{6 \times 7}{2} = \frac{7}{2} \qquad \ldots(*)$$

Remark. This does not mean that in a random throw of a dice, the player will get the number $(7/2) = 3·5$. In fact, one can never get this (fractional) number in a throw of a dice. Rather, this implies that if the player tosses the dice for a "long" period, then on the average toss he will get $(7/2) = 3·5$.

(b) The probability function of X (the sum of numbers obtained on two dice), is

Value of X : x	2	3	4	5	6	7	.....	11	12
Probability	1/36	2/36	3/36	4/36	5/36	6/36	.....	2/36	1/36

$$E(X) = \sum_i p_i x_i$$

$$= 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + 5 \times \frac{4}{36} + 6 \times \frac{5}{36} + 7 \times \frac{6}{36}$$

$$+ 8 \times \frac{5}{36} + 9 \times \frac{4}{36} + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36}$$

$$= \frac{1}{36}(2 + 6 + 12 + 20 + 30 + 42 + 40 + 36 + 30 + 22 + 12)$$

$$= \frac{1}{36} \times 252 = 7$$

Aliter. Let X_i be the number obtained on the ith dice ($i = 1, 2$) when thrown. Then the sum of the number of points on two dice is given by

$$S = X_1 + X_2$$

$$\Rightarrow \quad E(S) = E(X_1) + E(X_2) = \frac{7}{2} + \frac{7}{2} = 7 \qquad \text{[On using (*)]}$$

Remark. This result can be generalised to the sum of points obtained in a random throw of n dice. Then

$$E(S) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} (7/2) = \frac{7n}{2}$$

Example 6·3. *A box contains 2^n tickets among which ${}^nC_i$ tickets bear the number i ; $i = 0, 1, 2 \ldots, n$. A group of m tickets is drawn. What is the expectation of the sum of their numbers ?*

Solution. Let X_i ; $i = 1, 2, \ldots, m$ be the variable representing the number on the ith ticket drawn. Then the sum 'S' of the numbers on the tickets drawn is given by

$$S = X_1 + X_2 + \ldots + X_m = \sum_{i=1}^{m} X_i$$

$$\therefore \quad E(S) = \sum_{i=1}^{m} E(X_i)$$

Now X_i is a random variable which can take any one of the possible values 0, 1, 2, ..., n with respective probabilities.

$${}^nC_0/2^n,\ {}^nC_1/2^n,\ {}^nC_2/2^n,\ \ldots,\ {}^nC_n/2^n,$$

$$\therefore \quad E(X_i) = \frac{1}{2^n}\left[1.{}^nC_1 + 2.{}^nC_2 + 3.{}^nC_3 + \ldots + n.{}^nC_n\right]$$

$$= \frac{1}{2^n}\left[1.n + 2.\frac{n(n-1)}{2!} + 3.\frac{n(n-1)(n-2)}{3!} + \ldots + n.1\right]$$

$$= \frac{n}{2^n}\left[1 + (n-1) + \frac{(n-1)(n-2)}{2!} + \ldots + 1\right]$$

$$= \frac{n}{2^n}\left[{}^{n-1}C_0 + {}^{n-1}C_1 + {}^{n-1}C_2 + \ldots + {}^{n-1}C_{n-1}\right]$$

$$= \frac{n}{2^n} \cdot (1+1)^{n-1} = \frac{n}{2}$$

Hence $\quad E(S) = \sum_{i=1}^{m} (n/2) = \dfrac{m \cdot n}{2}$

Example 6·4. *In four tosses of a coin, let X be the number of heads. Tabulate the 16 possible outcomes with the corresponding values of X. By simple counting, derive the distribution of X and hence calculate the expected value of X.*

Solution. Let H represent a head, T a tail and X, the random variable denoting the number of heads.

S. No.	*Outcomes*	*No. of Heads (X)*	*S. No.*	*Outcomes*	*No. of Heads (X)*
1	*H H H H*	4	9	*H T H T*	2
2	*H H H T*	3	10	*T H T H*	2
3	*H H T H*	3	11	*T H H T*	2
4	*H T H H*	3	12	*H T T T*	1
5	*T H H H*	3	13	*T H T T*	1
6	*H H T T*	2	14	*T T H T*	1
7	*H T T H*	2	15	*T T T H*	1
8	*T T H H*	2	16	*T T T T*	0

The random variable X takes the values 0, 1, 2, 3 and 4. Since, from the above table, we find that the number of cases favourable to the coming of 0, 1, 2, 3 and 4 heads are 1, 4, 6, 4 and 1 respectively, we have

$$P(X=0) = \frac{1}{16}, \quad P(X=1) = \frac{4}{16} = \frac{1}{4}, \quad P(X=2) = \frac{6}{16} = \frac{3}{8},$$

$$P(X=3) = \frac{4}{16} = \frac{1}{4} \quad \text{and} \quad P(X=4) = \frac{1}{16}.$$

Thus the probability distribution of X can be summarised as follows :

x :	0	1	2	3	4
$p(x)$:	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

$$E(X) = \sum_{x=0}^{4} x\,p(x) = 1 \cdot \frac{1}{4} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{16}$$

$$= \frac{1}{4} + \frac{3}{4} + \frac{3}{4} + \frac{1}{4} = 2.$$

Example 6·5. *A coin is tossed until a head appears. What is the expectation of the number of tosses required ?* **[Delhi Univ. B.Sc., Oct. 1989]**

Solution. Let X denote the number of tosses required to get the first head. Then X can materialise in the following ways :

$$\therefore \qquad E(X) = \sum_{x=1}^{\infty} x\,p(x)$$

Event	x	Probability $p(x)$
H	1	1/2
TH	2	$1/2 \times 1/2 = 1/4$
TTH	3	$1/2 \times 1/2 \times 1/2 = 1/8$
$\vdots$	$\vdots$	$\vdots$

$$= 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 4 \times \frac{1}{16} + \dots \qquad \dots(*)$$

This is an arithmetic-geometric series with ratio of GP being $r = 1/2$.

Let $\qquad S = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + \dots$

Then $\qquad \frac{1}{2} S = \qquad \frac{1}{4} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{1}{16} + \dots$

$\therefore \quad (1 - \frac{1}{2}) S = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots$

$\Rightarrow \qquad \frac{1}{2} S = \dfrac{1/2}{1 - (1/2)} = 1$

[Since the sum of an infinite G.P. with first term a and common ratio r (< 1) is $a/(1 - r)$]

$\Rightarrow \qquad S = 2$

Hence, substituting in (*), we get

$$E(X) = 2$$

Example 6·6. *What is the expectation of the number of failures preceding the first success in an infinite series of independent trials with constant probability p of success in each trial ?* **[Delhi Univ. B.Sc., Oct. 1991]**

Solution. Let the random variable X denote the number of failures preceding the first success. Then X can take the values 0, 1, 2, ..., ∞. We have

$p(x) = P(X = x) = P[x$ failures precede the first success $] = q^x p$

where $q = 1 - p$ is the probability of failure in a trial. Then by def.

$$E(X) = \sum_{x=0}^{\infty} x\, p(x) = \sum_{x=0}^{\infty} x \cdot q^x p = pq \sum_{x=1}^{\infty} x\, q^{x-1}$$

$$= pq\,[1 + 2q + 3q^2 + 4q^3 + \dots]$$

Now $1 + 2q + 3q^2 + 4q^3 + \dots$ is an infinite arithmetic-geometric series.

Let $\qquad S = 1 + 2q + 3q^2 + 4q^3 + \dots$

$\qquad qS = \qquad q + 2q^2 + 3q^3 + \dots$

$\therefore \qquad (1 - q) S = 1 + q + q^2 + q^3 + \dots = \dfrac{1}{1 - q}$

$\Rightarrow \qquad S = \dfrac{1}{(1 - q)^2}$

$$\therefore \qquad 1 + 2q + 3q^2 + 4q^3 + \ldots = \frac{1}{(1-q)^2}$$

Hence $$E(X) = \frac{pq}{(1-q)^2} = \frac{pq}{p^2} = \frac{q}{p}$$

Example 6·7. *A box contains 'a' white and 'b' black balls. 'c' balls are drawn. Find the expected value of the number of white balls drawn.*

[Allahabad Univ. B.Sc., 1989; Indian Forest Service 1987]

Solution. Let a variable X_i, associated with ith draw, be defined as follows:

$X_i = 1$, if ith ball drawn is white

and $X_i = 0$, if ith ball drawn is black

Then the number 'S' of the white balls among 'c' balls drawn is given by

$$S = X_1 + X_2 + \ldots + X_c = \sum_{i=1}^{c} X_i \quad \Rightarrow \quad E(S) = \sum_{i=1}^{c} E(X_i)$$

Now $$P(X_i = 1) = P(\text{of drawing a white ball}) = \frac{a}{a+b}$$

and $$P(X_i = 0) = P(\text{of drawing a black ball}) = \frac{b}{a+b}$$

$$\therefore \qquad E(X_i) = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) = \frac{a}{a+b}$$

Hence $$E(S) = \sum_{i=1}^{c} \left(\frac{a}{a+b}\right) = \frac{ca}{a+b}$$

Example 6·8. *Let variate X have the distribution*

$$P(X=0) = P(X=2) = p\,;\; P(X=1) = 1-2p, \quad \text{for } 0 \le p \le \tfrac{1}{2}.$$

For what p is the Var (X) a maximum ?

[Delhi Univ. B.Sc. (Maths Hons.) 1987, 85]

Solution. Here the r.v. X takes the values 0, 1 and 2 with respective probabilities p, $1 - 2p$ and p, $0 \le p \le \frac{1}{2}$.

$$\therefore \qquad E(X) = 0 \times p + 1 \times (1-2p) + 2 \times p = 1$$

$$E(X^2) = 0 \times p + 1^2 \times (1-2p) + 2^2 \times p = 1 + 2p$$

$$\therefore \qquad Var(X) = E(X^2) - [E(X)]^2 = 2p\;;\; 0 \le p \le \tfrac{1}{2}$$

Obviously $Var(X)$ is maximum when $p = \frac{1}{2}$, and

$$[Var(X)]_{max} = 2 \times \tfrac{1}{2} = 1$$

Example 6·9. $Var(X) = 0 \quad \Rightarrow \quad P[X = E(X)] = 1$. Comment.

Solution. $Var(X) = E[X - E(X)]^2 = 0$

$\Rightarrow \qquad [X - E(X)]^2 = 0$, with probability 1

$\Rightarrow \qquad [X - E(X)] = 0$, with probability 1

$\Rightarrow \qquad P[X = E(X)] = 1$

Example 6·10. *Explain by means of an example that a probability distribution is not uniquely determined by its moments.*

Solution. Consider a r.v. X with p.d.f. [*c.f.* Log-Normal distribution § 8·2·15

$$f(x) = \frac{1}{\sqrt{2\pi}\,x} \cdot \exp\left[-\tfrac{1}{2}(\log x)^2\right] \; ; \; x > 0 \qquad \ldots(*)$$
$$= 0 \; ; \quad \text{otherwise}$$

Consider, another r. v. Y with p.d.f.

$$g(y) = \Big[1 + a \sin(2\pi \log y)\Big] f(y) = g_a(y), \text{(say)}, \quad y > 0 \quad \ldots(**)$$

which, for $-1 \le a \le 1$, represents a family of probability distributions.

$$E(Y^r) = \int_0^\infty y^r \left\{1 + a \sin(2\pi \log y)\right\} f(y)\, dy$$

$$= \int_0^\infty y^r f(y)\, dy + a \cdot \int_0^\infty y^r \cdot \sin(2\pi \log y) f(y)\, dy.$$

$$= EX^r + a \cdot \frac{1}{\sqrt{2\pi}} \int_0^\infty y^r \cdot \sin(2\pi \log y) \cdot \frac{1}{y} \exp\left[-\tfrac{1}{2}(\log y)^2\right] dy$$

$$= EX^r + \frac{a}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{rz - z^2/2} \cdot \sin(2\pi z)\, dz$$

$$\left[\log y = z \;\Rightarrow\; y = e^z\right]$$

$$= EX^r + \frac{a \cdot e^{r^2/2}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(z-r)^2} \cdot \sin(2\pi z)\, dz$$

$$= EX^r + \frac{a \cdot e^{r^2/2}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-y^2/2} \cdot \sin(2\pi y)\, dy$$

[$z - r = y \;\Rightarrow\; \sin(2\pi z) = \sin(2\pi r + 2\pi y) = \sin 2\pi y$, r being a positive integer].

$$= EX^r,$$

the value of the integral being zero, since the integrand is an odd function of y.

$\Rightarrow$ $E(Y^r)$ is independent of 'a' in (**).

Hence, $\{\, g(y) = g_a(y); \; -1 \le a \le 1 \,\}$, represents a family of distributions, each different from the other, but having the same moments. This explains that the moments may not determine a distribution uniquely.

Example 6·11. *Starting from the origin, unit steps are taken to the right with probability p and to the left with probability q (= 1 − p). Assuming independent movements, find the mean and variance of the distance moved from origin after n steps (Random Walk Problem).*

Solution. Let us associate a variable X_i with the ith step defined as follows :

$X_i = +1$, if the ith step is towards the right,

$= -1$, if the ith step is towards the left.

Then $S = X_1 + X_2 + \ldots + X_n = \Sigma X_i$, represents the random distance moved from origin after n steps.

$$E(X_i) = 1 \times p + (-1) \times q = p - q$$

$$E(X_i^2) = 1^2 \times p + (-1)^2 \times q = p + q = 1$$

$$\therefore \quad Var(X_i) = E(X_i^2) - [E(X_i)]^2 = (q+p)^2 - (p-q)^2 = 4pq$$

$$\therefore \quad E(S_n) = \sum_{i=1}^{n} E(X_i) = n(p-q)$$

$$V(S_n) = \sum_{i=1}^{n} V(X_i) = 4npq$$

[$\because$ *Movements of steps are independent*].

Example 6·12. *Let r.v. X have a density function $f(\cdot)$, cumulative distribution function $F(\cdot)$, mean μ and variance σ^2. Define $Y = \alpha + \beta X$, where α and β are constants satisfying $-\infty < \alpha < \infty$* and $\beta > 0$.

(a) Select α and *β so that Y has mean 0 and variance 1.*

(b) What is the correlation coefficient ρ_{XY} between X and Y ?

(c) Find the cumulative distribution function of Y in terms of α, β and $F(\cdot)$.

(d) If X is symmetrically distributed about μ, is Y necessarily symmetrically distributed about its mean?

Solution. *(a)* $E(X) = \mu$, $Var(X) = \sigma^2$. We want α and β s.t.

$$E(Y) = E(\alpha + \beta X) = \alpha + \beta\mu = 0 \qquad \ldots(1)$$

$$Var(Y) = Var(\alpha + \beta X) = \beta^2 . \sigma^2 = 1 \qquad \ldots(2)$$

Solving *(1)* and *(2)* we get :

$$\beta = 1/\sigma, \ (>0) \text{ and } \alpha = -\mu/\sigma \qquad \ldots(3)$$

(b) $\quad Cov(X, Y) = E(XY) - E(X)E(Y) = E[X(\alpha + \beta X)]$

[$\because E(Y) = 0$]

$$= \alpha . E(X) + \beta . E(X^2) = \alpha\mu + \beta[\sigma^2 + \mu^2]$$

$$\therefore \quad \rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{\alpha\mu + \beta[\sigma^2 + \mu^2]}{\sigma . 1} \qquad (\because \sigma_Y = 1)$$

$$= \frac{1}{\sigma^2}\left[-\mu^2 + \sigma^2 + \mu^2\right] = 1 \qquad \text{[On using (3)]}$$

(c) Distribution function $G_Y(\cdot)$ of Y is given by :

$$G_Y(y) = P(Y \le y) = P[\alpha + \beta X \le y]$$

$$= P(X \le (y - \alpha)/\beta)$$

$$\Rightarrow \quad G_Y(y) = F_X\left(\frac{y-\alpha}{\beta}\right);$$

(d) We have : $\quad Y = \alpha + \beta X = \dfrac{1}{\sigma}(X - \mu) = \beta(X - \mu)$ $\qquad$ [*On using (3)*]

Since X is given to be symmetrically distributed about mean μ, $(X - \mu)$ and $-(X - \mu)$ have the same distribution.

Hence $Y = \beta(X - \mu)$ and $-Y = -\beta(X - \mu)$ have the same distribution. Since $E(Y) = 0$, we conclude that Y is symmetrically distributed about its mean.

Example 6·13. *Let X be a r.v. with mean μ and variance σ^2. Show that $E(X-b)^2$, as a function of b, is minimised when $b = \mu$.*

Solution.
$$\begin{aligned} E(X-b)^2 &= E[(X-\mu)+(\mu-b)]^2 \\ &= E(X-\mu)^2 + (\mu-b)^2 + 2(\mu-b)E(X-\mu) \\ &= Var(X) + (\mu-b)^2 \qquad [\because E(X-\mu)=0] \end{aligned}$$

$$\Rightarrow \quad E(X-b)^2 \geq Var(X), \qquad \ldots(*)$$

since $(\mu - b)^2$, being the square of a real quantity is always non-negative.

The sign of equality holds in (*) iff

$$(\mu - b)^2 = 0 \quad \Rightarrow \quad \mu = b.$$

Hence $E(X-b)^2$ is minimised when $\mu = b$ and its minimum value is $E(X-\mu)^2 = \sigma_X{}^2$.

Remark. This result states that the sum of squares of deviations is minimum when taken about mean.

[Also see § 2·4, Property 3 of Arithmetic Mean]

Example 6·14. *Let $a_1, a_2, \ldots, a_n$ be arbitrary real numbers and $A_1, A_2, \ldots, A_n$ be events. Prove that*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j P(A_i A_j) \geq 0$$

[Delhi Univ. B.A. (Spl. Course – Stat. Hons.), 1986]

Solution. Let us define the indicator variable :

$$X_i = I_{A_i} = 1 \quad \text{if } A_i \text{ occurs}$$
$$= 0 \quad \text{if } \bar{A}_i \text{ occurs}.$$

Then using (6·2b) :

$$E(X_i) = P(A_i)\,; \quad (i = 1, 2, \ldots, n) \qquad \ldots(i)$$

Also $\quad X_i X_j = I_{A_i \cap A_j}$

$$\Rightarrow \quad E(X_i X_j) = P(A_i A_j) \qquad \ldots(ii)$$

Consider, for real numbers $a_1, a_2, \ldots, a_n$, the expression $\left(\sum_{i=1}^{n} a_i X_i\right)^2$, which is always non-negative.

$$\Rightarrow \quad \left(\sum_{i=1}^{n} a_i X_i\right)^2 \geq 0$$

$$\Rightarrow \quad \left(\sum_{i=1}^{n} a_i X_i\right)\left(\sum_{j=1}^{n} a_j X_j\right) \geq 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j X_i X_j \geq 0, \qquad \ldots(iii)$$

for all a_i 's and a_j 's.

Since expected value of a non-negative quantity is always non-negative, on taking expectations of both sides in *(iii)* and using *(i)* and *(ii)* we get :

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \; E(X_i X_j) \geq 0 \;\Rightarrow\; \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \; P(A_i A_j) \geq 0.$$

Example 6·15. *In a sequence of Bernoulli trials, let X be the length of the run of either successes or failures starting with the first trial. Find E (X) and V (X).*

Solution. Let 'p' denote the probability of success. Then $q = 1 - p$ is the probability of faiuure. $X = 1$ means that we can have any one of the possibilities SF and FS with respective probabilities pq and qp.

$$\therefore \quad P(X = 1) = P(SF) + P(FS) = pq + qp = 2pq$$

Similarly

$$P(X = 2) = P(SSF) + P(FFS) = p^2q + q^2p$$

In general

$$P(X = r) = P[SSS...SF] + P[FFF...FS] = p^r.q + q^r.p$$

$$\therefore \quad E(X) = \sum_{r=1}^{\infty} r\,P(X = r) = \sum_{r=1}^{\infty} r\,(p^r.q + q^r.p)$$

$$= pq\;\Big[\sum_{r=1}^{\infty} r.p^{r-1} + \sum_{r=1}^{\infty} r.q^{r-1}\Big]$$

$$= pq\;[(1 + 2p + 3p^2 + ...) + (1 + 2q + 3q^2 + ...)]$$

$$= pq\;[(1-p)^{-2} + (1-q)^{-2}] = pq\;[q^{-2} + p^{-2}]$$

(See Remark to Example 6·17)

$$= pq\left[\frac{1}{q^2} + \frac{1}{p^2}\right] = \frac{p}{q} + \frac{q}{p}.$$

$$V(X) = E(X^2) - \{E(X)\}^2 = E\{X(X-1)\} + E(X) - \{E(X)\}^2$$

Now

$$E\{X(X-1)\} = \sum_{r=2}^{\infty} r(r-1)\,P(X = r) = \sum_{r=2}^{\infty} r(r-1)\,(p^rq + q^rp)$$

$$= \sum_{r=2}^{\infty} r(r-1)\,p^rq + \sum_{r=2}^{\infty} r(r-1)\,q^rp$$

$$= p^2q \sum_{r=2}^{\infty} r(r-1)\,p^{r-2} + q^2p \sum_{r=2}^{\infty} r(r-1)\,q^{r-2}$$

$$= 2p^2q \sum_{r=2}^{\infty} \frac{r(r-1)}{2}\,p^{r-2} + 2q^2p \sum_{r=2}^{\infty} \frac{r(r-1)}{2}\,q^{r-2}$$

$$= 2p^2q\,(1-p)^{-3} + 2q^2p\,(1-q)^{-3}$$

$$= 2\left(\frac{p^2}{q^2} + \frac{q^2}{p^2}\right)$$

$$V(X) = 2\left(\frac{p^2}{q^2} + \frac{q^2}{p^2}\right) + \left(\frac{p}{q} + \frac{q}{p}\right) - \left(\frac{p}{q} + \frac{q}{p}\right)^2$$

$$= \left(\frac{p}{q} - \frac{q}{p}\right)^2 + \left(\frac{p}{q} + \frac{q}{p}\right)$$

Aliter. Proceed as in Example 6·17.

Example 6·16. *A deck of n numbered cards is thoroughly shuffled and the cards are inserted into n numbered cells one by one. If the card number 'i' falls in the cell 'i', we count it as a match, otherwise not. Find the mean and variance of total number of such matches.* **[Delhi Univ. B.Sc., (Stat. Hons.), 1988]**

Solution. Let us associate a random variable, X_i with the ith draw defined as follows :

$$X_i = \begin{cases} 1, & \text{if the } i\text{th card dealt has the number '}i\text{' on it} \\ 0, & \text{otherwise} \end{cases}$$

Then the total number of matches 'S' is given by

$$S = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

$$\therefore \quad E(S) = \sum_{i=1}^{n} E(X_i)$$

Now $\quad E(X_i) = 1.P(X_i = 1) + 0.P(X_i = 0) = P(X_i = 1) = \dfrac{1}{n}$

Hence $\quad E(S) = \sum_{i=1}^{n}\left(\dfrac{1}{n}\right) = n \cdot \dfrac{1}{n} = 1$

$$V(S) = V(X_1 + X_2 + \ldots + X_n)$$

$$= \sum_{i=1}^{n} V(X_i) + 2 \sum\sum_{\substack{i,j=1 \\ i<j}}^{n} \text{Cov}(X_i, X_j) \quad \ldots(1)$$

Now $\quad V(X_i) = E(X_i^2) - \{E(X_i)\}^2$

$$= 1^2 . P(X_i = 1) + 0^2 . P(X_i = 0) - \left(\frac{1}{n}\right)^2$$

$$= \frac{1}{n} - \frac{1}{n^2} = \frac{n-1}{n^2} \quad \ldots(2)$$

$$\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)\, E(X_j) \quad \ldots(3)$$

$$E(X_i X_j) = 1.P(X_i X_j = 1) + 0.P(X_i X_j = 0)$$

$$= \frac{(n-2)!}{n!} = \frac{1}{n(n-1)},$$

since $X_i X_j = 1$ if and only if both card numbers i and j are in their respective matching places and there are $(n-2)!$ arrangements of the remaining cards that correspond to this event.

Substituting in (3), we get

$$\text{Cov}(X_i, X_j) = \frac{1}{n(n-1)} - \frac{1}{n}\cdot\frac{1}{n} = \frac{1}{n^2(n-1)} \qquad ...(4)$$

Substituting from (2) and (4) in (1), we have

$$V(S) = \sum_{i=1}^{n}\left(\frac{n-1}{n^2}\right) + 2\sum_{\substack{i=1\\i<j}}^{n}\sum_{j=1}^{n}\left[\frac{1}{n^2(n-1)}\right]$$

$$= n\left(\frac{n-1}{n^2}\right) + 2.{}^nC_2\,\frac{1}{n^2(n-1)} = \frac{n-1}{n} + \frac{1}{n} = 1$$

Example 6·17. *If t is any positive real number, show that the function defined by*

$$p(x) = e^{-t}(1-e^{-t})^{x-1} \qquad ...(*)$$

can represent a probability function of a random variable X assuming the values $1, 2, 3, \ldots$ Find the $E(X)$ and $Var(X)$ of the distribution.

[Nagpur Univ. B.Sc., 1988]

Solution. We have

$$e^t > 1,\ \forall\ t > 0 \Rightarrow e^{-t} < 1 \Rightarrow 1 - e^{-t} > 0$$

Also $\qquad e^{-t} = \dfrac{1}{e^t} > 0,\ \forall\ t > 0$

Hence $\qquad p(x) = e^{-t}(1-e^{-t})^{x-1} \geq 0\ \forall\ t > 0, x = 1, 2, 3, \ldots$

Also $\qquad \sum_{x=1}^{\infty} p(x) = e^{-t}\sum_{x=1}^{\infty}(1-e^{-t})^{x-1} = e^{-t}\sum_{x=1}^{\infty} a^{x-1};\qquad [a = 1-e^{-t}]$

$$= e^{-t}(1 + a + a^2 + a^3 + \ldots) = e^{-t}\times\frac{1}{(1-a)}$$

$$= e^{-t}[1-(1-e^{-t})]^{-1} = e^{-t}(e^{-t})^{-1} = 1$$

Hence $p(x)$ defined in (*) represents the probability function of a r.v. X.

$$E(X) = \Sigma x.p(x) = e^{-t}\sum_{x=1}^{\infty} x(1-e^{-t})^{x-1}$$

$$= e^{-t}\sum_{x=1}^{\infty} x.a^{x-1};\ [a = 1-e^{-t}]$$

$$= e^{-t}(1 + 2a + 3a^2 + 4a^3 + \ldots) = e^{-t}(1-a)^{-2} \qquad ...(*)$$

$$= e^{-t}(e^{-t})^{-2} = e^t$$

$$E(X^2) = \Sigma x^2 p(x) = e^{-t}\sum_{x=1}^{\infty} x^2.a^{x-1}$$

$$= e^{-t}[1 + 4a + 9a^2 + 16a^3 + \ldots]$$

$$= e^{-t}(1+a)(1-a)^{-3} = e^{-t}(2-e^{-t})e^{3t}$$

Hence $\qquad Var(X) = E(X^2) - [E(X)]^2 = e^{-t}(2-e^{-t})e^{3t} - e^{2t}$

$$= e^{2t}[(2-e^{-t})-1] = e^{2t}(1-e^{-t})$$

$$= e^t(e^t - 1)$$

Remark.

(*i*) Consider $\qquad S = 1 + 2a + 3a^2 + 4a^3 + \ldots$ (Arithmetico-geometric series)

$$\Rightarrow \qquad aS = a + 2a^2 + 3a^3 + \ldots$$

$$\Rightarrow \qquad (1-a)S = 1 + a + a^2 + a^3 + \ldots = 1/(1-a) \Rightarrow S = (1-a)^{-2}$$

$$\sum_{x=1}^{\infty} x\, a^{x-1} = 1 + 2a + 3a^2 + 4a^3 + \ldots = (1-a)^{-2} \qquad \ldots(*)$$

(ii) Consider

$$S = 1 + 2^2 . a + 3^2 . a^2 + 4^2 . a^3 + 5^2 . a^4 + \ldots$$

$$\Rightarrow \qquad S = 1 + 4a + 9a^2 + 16a^3 + 25a^4 + \ldots$$

$$-3aS = -3a - 12a^2 - 27a^3 - 48a^4 - \ldots$$

$$+3a^2S = +3a^2 + 12a^3 + 27a^4 + \ldots$$

$$-a^3S = -a^3 - 4a^4 - \ldots$$

Adding the above equations we get :

$$(1-a)^3 S = 1 + a \qquad \Rightarrow \qquad S = (1+a)(1-a)^{-3} \qquad \ldots(**)$$

$$\sum_{x=1}^{\infty} x^2 a^{x-1} = 1 + 4a + 9a^2 + 16a^3 + \ldots = (1+a)(1-a)^{-3}$$

The results in (*) and (**) are quite useful for numerical problems and should be committed to memory.

Example 6·18. *A man with n keys wants to open his door and tries the keys independently and at random. Find the mean and variance of the number of trials required to open the door (i) if unsuccessful keys are not eliminated from further selection, and (ii) if they are.* **[Rajasthan Univ. B.Sc.(Hons.), 1992]**

Solution. *(i)* Suppose the man gets the first success at the xth trial, *i.e.*, he is unable to open the door in the first $(x-1)$ trials. If unsuccessful keys are not eliminated then X is a random variable which can take the values 1, 2, 3,.... ad infinity.

Probability of success at the first trial $= 1/n$

$\therefore$ Probability of failure at the first trial $= 1 - (1/n)$

If unsuccessful keys are not eliminated then the probability of success and consequently of failure is constant for each trial.

Hence $\quad p(x) =$ Probability of 1st success at the xth trial

$$= \left(1 - \frac{1}{n}\right)^{x-1} . \frac{1}{n}$$

$$\therefore \qquad E(X) = \sum_{x=1}^{\infty} x\, p(x) = \sum_{x=1}^{\infty} x \left(1 - \frac{1}{n}\right)^{x-1} . \frac{1}{n}$$

$$= \frac{1}{n} \sum_{x=1}^{\infty} x A^{x-1}, \text{ where } A = 1 - \frac{1}{n}$$

$$E(X) = \frac{1}{n}[1 + 2A + 3A^2 + 4A^3 + \ldots] = \frac{1}{n}(1-A)^{-2}$$

[See (*), Example (6·17)]

$$\frac{1}{n}\left[1-\left(1-\frac{1}{n}\right)\right]^{-2} = n$$

$$E(X^2) = \sum_{x=1}^{\infty} x^2 p(x) = \sum_{x=1}^{\infty} x^2 \left(1-\frac{1}{n}\right)^{x-1} \cdot \frac{1}{n}$$

$$= \frac{1}{n} \sum_{x=1}^{\infty} x^2 A^{x-1}$$

$$= \frac{1}{n}[1 + 2^2 . A + 3^2 . A^2 + 4^2 . A^3 + \ldots]$$

$$= \frac{1}{n}(1+A)(1-A)^{-3} \qquad \text{[See (**), Example (6·17)]}$$

$$= \frac{1}{n}\left[1+\left(1-\frac{1}{n}\right)\right]\left[1-\left(1-\frac{1}{n}\right)\right]^{-3}$$

$$= (2n-1)\,n$$

Hence $\qquad V(X) = E(X^2) - [E(X)]^2 = (2n-1)\,n - n^2 = n^2 - n = n(n-1)$

(ii) If unsuccessful keys are eliminated from further selection, then the random variable X will take the values from 1 to n. In this case, we have

Probability of success at the first trial = $1/n$

Probability of success at the 2nd trial = $1/(n-1)$

Probability of success at the 3rd trial = $1/(n-2)$

and so on.

Hence probability of 1st succes at 2nd trial = $\left(1-\frac{1}{n}\right)\frac{1}{n-1} = \frac{1}{n}$

Probability of first success at the third trial

$$= \left(1-\frac{1}{n}\right)\left(1-\frac{1}{n-1}\right)\cdot\frac{1}{n-2} = \frac{1}{n},$$

and so on. In general, we have

$$p(x) = \text{Probability of first success at the } x\text{th trial} = \frac{1}{n}$$

$$\therefore \qquad E(X) = \sum_{x=1}^{\infty} x\,p(x) = \frac{1}{n}\sum_{x=1}^{n} x = \frac{n+1}{2}$$

$$E(X^2) = \sum_{x=1}^{n} x^2 p(x) = \frac{1}{n}\sum_{x=1}^{n} x^2 = \frac{(n+1)(2n+1)}{6}$$

Hence $\quad V(X) = E(X^2) - [E(X)]^2 = \dfrac{(n+1)(2n+1)}{6} - \left(\dfrac{n+1}{2}\right)^2$

$$= \frac{n+1}{12}\Big[\,2(2n+1) - 3(n+1)\,\Big] = \frac{n^2-1}{12}$$

Example 6·19. *In a lottery m tickets are drawn at a time out of n tickets numbered 1 to n. Find the expectation and the variance of the sum S of the numbers on the tickets drawn.* **[Delhi Univ. B.Sc. (Maths Hons.), 1987]**

Solution. Let X_i denote the score on the ith ticket drawn.

Then $$S = X_1 + X_2 + \dots + X_m = \sum_{i=1}^{m} X_i,$$

is the total score on the m tickets drawn.

$$\therefore \quad E(S) = \sum_{i=1}^{m} E(X_i)$$

Now each X_i is a random variable which assumes the values 1, 2, 3, ..., n each with equal probability $1/n$.

$$\therefore \quad E(X_i) = \frac{1}{n}(1 + 2 + 3 + \dots + n) = \frac{(n+1)}{2}$$

Hence $$E(S) = \sum_{i=1}^{m}\left(\frac{n+1}{2}\right) = \frac{m(n+1)}{2}$$

$$V(S) = V(X_1 + X_2 + \dots + X_m)$$

$$= \sum_{i=1}^{m} V(X_i) + 2 \sum_{\substack{i,j \\ i<j}}\sum \mathrm{Cov}(X_i, X_j)$$

$$E(X_i^2) = \frac{1}{n}\left(1^2 + 2^2 + 3^2 + \dots + n^2\right)$$

$$= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

$$\therefore \quad V(X_i) = E(X_i^2) - [E(X_i)]^2$$

$$= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

Also $$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)\,E(X_j)$$

To find $E(X_i X_j)$ we note that the variables X_i and X_j can take the values as shown below :

X_i	X_j
1	2, 3, ..., n
2	1, 3, ..., n
⋮	⋮
n	1, 2, ..., $(n-1)$

Thus the variable $X_i X_j$ can take $n(n-1)$ possible values and $P(X_i = l \cap X_j = k) = \dfrac{1}{n(n-1)}$, $k \neq l$. Hence

$$E(X_i X_j) = \frac{1}{n(n-1)} \begin{bmatrix} 1.2 + 1.3 + \ldots\ldots + 1.n \\ + 2.1 + 2.3 + \ldots\ldots + 2.n \\ + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ + n.1 + n.2 + \ldots + n.(n-1) \end{bmatrix}$$

$$= \frac{1}{n(n-1)} \begin{bmatrix} 1(1+2+3+\ldots\ldots+n) - 1^2 \\ + 2(1+2+3+\ldots\ldots+n) - 2^2 \\ + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ + n(1+2+\ldots\ldots\overline{n-1}+n) - n^2 \end{bmatrix}$$

$$= \frac{1}{n(n-1)} \left[(1+2+3+\ldots+n)^2 - (1^2+2^2+\ldots+n^2) \right]$$

$$= \frac{1}{n(n-1)} \left[\left\{ \frac{n(n+1)}{2} \right\}^2 - \frac{n(n+1)(2n+1)}{6} \right]$$

$$= \frac{(n+1)(3n^2-n-2)}{12(n-1)}$$

$$\therefore \quad \text{Cov}(X_i, X_j) = \frac{(n+1)(3n^2-n-2)}{12(n-1)} - \left(\frac{n+1}{2} \right)^2$$

$$= \frac{(n+1)}{12(n-1)} \left[3n^2 - n - 2 - 3(n^2-1) \right]$$

$$= -\frac{(n+1)}{12}$$

Hence
$$V(S) = \sum_{i=1}^{m} \left(\frac{n^2-1}{12} \right) + 2 \sum_{i<j=1}^{m}\sum \left\{ -\frac{(n+1)}{12} \right\}$$

$$= \frac{m(n^2-1)}{12} + 2 \cdot \frac{m(m-1)}{2!} \left\{ \frac{-(n+1)}{12} \right\},$$

[since there are ${}^mC_2$ covariance terms in $\text{Cov}(X_i, X_j)$]

$$\therefore \quad V(S) = \frac{m(n+1)}{12} \Big[(n-1) - (m-1) \Big] = \frac{m(n+1)(n-m)}{12}$$

Example 6·20. *A die is thrown $(n+2)$ times. After each throw a '+' is recorded for 4, 5 or 6 and '–' for 1, 2 or 3, the signs forming an ordered sequence. To each, except the first and the last sign, is attached a characteristic random variable which takes the value 1 if both the neighbouring signs differ from the one between them and 0 otherwise. If $X_1, X_2, \ldots, X_n$ are characteristic random variables, find the mean and variance of $X = \sum_{i=1}^{n} X_i$.*

Solution. $\quad X = \sum_{i=1}^{n} X_i \quad \Rightarrow \quad E(X) = \sum_{i=1}^{n} E(X_i)$

Now $\quad E(X_i) = 1.P(X_i = 1) + 0.P(X_i = 0) = P(X_i = 1)$

For $X_i = 1$, there are the following two mutually exclusive possibilities :

(i) $\quad - \quad + \quad -$, $\qquad$ (ii) $\quad + \quad - \quad +$

and since the probability of each sign is $\frac{1}{2}$, we have by addition probability theorem:

$$P(X_i = 1) = P(i) + P(ii) = \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

$$\therefore \quad E(X_i) = \frac{1}{4}$$

Hence $\quad E(X) = \sum_{i=1}^{n}\left(\frac{1}{4}\right) = \frac{n}{4},$

$$V(X) = V(X_1 + V_2 + \ldots + X_n)$$

$$= \sum_{i=1}^{n} V(X_i) + 2 \sum\sum_{i<j} \text{Cov}(X_i, X_j) \qquad \ldots(*)$$

Now $\quad E(X_i^2) = 1^2.P(X_i = 1) + 0^2.P(X_i = 0) = P(X_i = 1) = \frac{1}{4}$

$$\therefore \quad V(X_i) = E(X_i^2) - [E(X_i)]^2 = \frac{1}{4} - \frac{1}{16} = \frac{3}{16}$$

Now $\quad E(X_i X_j) = 1.P(X_i = 1 \cap X_j = 1) + 0.P(X_i = 0 \cap X_j = 0)$

$$4 + 0.P(X_i = 1 \cap X_j = 0) + 0.P(X_i = 0 \cap X_j = 1)$$

$$= P(X_i = 1 \cap X_j = 1)$$

Since there are the following two mutually exclusive possibilities for the event : $(X_i = 1 \cap X_j = 1)$,

(i) $- \quad + \quad - \quad +$

(ii) $+ \quad - \quad + \quad -$, we have

$$P(X_i = 1 \cap X_j = 1) = P(i) + P(ii) = \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^4 = \frac{1}{8}$$

$$\therefore \quad \text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)\,E(X_j)$$

$$= \frac{1}{8} - \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

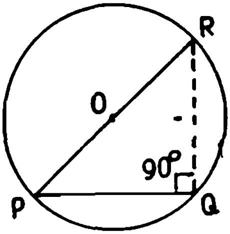
Hence $\quad V(X) = \sum_{i=1}^{n} (3/16) + 2 \sum\sum_{i<j} \text{Cov}(X_i, X_j) \qquad$ [From (*)]

$$= \frac{3n}{16} + 2\,[\text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_3) + \ldots + \text{Cov}(X_{n-1}, X_n)]$$

$$= \frac{3n}{16} + 2(n-1)\,.\,\frac{1}{16} = \frac{5n-2}{16}$$

Example 6·21. *From a point on the circumference of a circle of radius 'a', a chord is drawn in a random direction, (all directions are equally likely). Show that the expected value of the length of the chord is $4a/\pi$ and that the variance of the* *length is* $2a^2\,(1 - 8/\pi^2)$. *Also show that the chance is 1/3 that the length of the chord will exceed the length of the side of an equilateral triangle inscribed in the circle.*

Solution. Let P be any point on the circumference of a circle of radius 'a' and centre 'O'. Let PQ be any chord drawn at random and let $\angle OPQ = \theta$. Obviously, θ ranges from $-\pi/2$ to $\pi/2$. Since all the directions are equally likely, the probability differential of θ is given by the rectangular distribution (*c.f.* Chapter 8) :



$$dF(\theta) = f(\theta)\,d(\theta) = \frac{d\theta}{\pi/2 - (-\pi/2)}$$

$$= \frac{d\theta}{\pi}, \quad \frac{-\pi}{2} \le \theta \le \frac{\pi}{2}$$

Now, since $\angle PQR$ is a right angle, (angle in a semi-circle), we have

$$\frac{PQ}{PR} = \cos\theta \;\Rightarrow\; PQ = PR\cos\theta = 2a\cos\theta$$

$$E(PQ) = \int_{-\pi/2}^{\pi/2} (PQ)\,f(\theta)\,d(\theta) = \frac{2a}{\pi}\int_{-\pi/2}^{\pi/2} \cos\theta\,d\theta$$

$$= \frac{2a}{\pi}\Big|\sin\theta\Big|_{-\pi/2}^{\pi/2} = \frac{4a}{\pi}$$

$$E[(PQ)^2] = \int_{-\pi/2}^{\pi/2} [PQ]^2 f(\theta)\,d\theta = \frac{4a^2}{\pi}\int_{-\pi/2}^{\pi/2} \cos^2\theta\,d\theta$$

$$= \frac{4a^2}{\pi}\int_0^{\pi/2} 2\cos^2\theta\,d\theta = \frac{4a^2}{\pi}\int_0^{\pi/2} (1 + \cos 2\theta)\,d\theta$$

(Since $\cos^2\theta$ is an even function of θ)

$$= \frac{4a^2}{\pi}\Big|\theta + \frac{\sin 2\theta}{2}\Big|_0^{\pi/2} = \frac{4a^2}{\pi}\cdot\frac{\pi}{2} = 2a^2$$

$$\therefore \quad V(PQ) = E[(PQ)^2] - [E(PQ)]^2 = 2a^2 - \frac{16a^2}{\pi^2} = 2a^2\left(1 - \frac{8}{\pi^2}\right)$$

We know that the length of the side of an equilateral triangle inscribed in a circle of radius 'a' is $a\sqrt{3}$. Hence

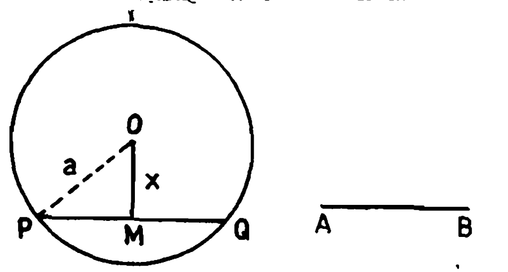
$$P(PQ > a\sqrt{3}) = P(2a\cos\theta > a\sqrt{3}) = P\left(\cos\theta > \frac{\sqrt{3}}{2}\right)$$

$$= P\left(|\theta| < \frac{\pi}{6}\right) = P\left(\frac{-\pi}{6} < \theta < \frac{\pi}{6}\right)$$

$$= \int_{-\pi/6}^{\pi/6} f(\theta)\,d\theta = \frac{1}{\pi}\int_{-\pi/6}^{\pi/6} 1\,.\,d\theta = \frac{1}{\pi}\cdot\frac{\pi}{3} = \frac{1}{3}$$

Example 6·22. *A chord of a circle of radius 'a' is drawn parallel to a given straight line, all distances from the centre of the circle being equally likely. Show that the expected value of the length of the chord is $\pi a/2$ and that the variance of the length is $a^2\,(32 - 3\pi^2)/12$. Also show that the chance is 1/2 that the length of* *the chord will exceed the length of the side of an equilateral triangle inscribed in the circle.*

Solution. Let PQ be the chord of a circle with centre O and radius 'a' drawn at random parallel to the given straight line AB. Draw $OM \perp PQ$. Let $OM = x$. Obviously x ranges from $-a$ to a. Since all distances from the centre are equally



likely, the probability that a random value of x will lie in the small interval 'dx' is given by the rectangular distribution [c.f. Chapter 8]:

$$dF(x) = f(x)\,dx = \frac{dx}{a-(-a)} = \frac{dx}{2a}, \quad -a \le x \le a$$

Length of the chord is

$$PQ = 2PM = 2\sqrt{a^2 - x^2}$$

Hence
$$E(PQ) = \int_{-a}^{a} PQ\, dF(x) = \frac{2}{2a}\int_{-a}^{a} \sqrt{a^2 - x^2}\, dx$$

$$= \frac{2}{a}\int_{0}^{a} \sqrt{a^2 - x^2}\, dx,$$

(since integrand is an even function of x).

$$= \frac{2}{a}\left| \frac{1}{2}x\sqrt{a^2 - x^2} + \frac{1}{2}a^2 \sin^{-1}\left(\frac{x}{a}\right)\right|_0^a$$

$$= \frac{2}{a}\cdot\frac{a^2}{2}\cdot\frac{\pi}{2} = \frac{\pi a}{2}$$

$$E[(PQ)^2] = \int_{-a}^{a} (PQ)^2\, dF(x) = \frac{4}{2a}\int_{-a}^{a} (a^2 - x^2)\, dx$$

$$= \frac{4}{a}\int_{0}^{a} (a^2 - x^2)\, dx = \frac{4}{a}\left| a^2 x - \frac{x^3}{3}\right|_0^a$$

$$= \frac{4}{a}\cdot\frac{2a^3}{3} = \frac{8a^2}{3}$$

Hence

$$\text{Var (length of chord)} = E[(PQ)^2] - [E(PQ)]^2 = \frac{8a^2}{3} - \frac{\pi^2 a^2}{4}$$

$$= \frac{a^2}{12}(32 - 3\pi^2)$$

The length of the chord is greater than the side of the equilateral triangle inscribed in the circle if

$$2\sqrt{a^2-x^2} > a\sqrt{3} \quad \Rightarrow \quad 4(a^2-x^2) > 3a^2$$

i.e.,

$$x^2 < a^2/4 \quad \Rightarrow \quad |x| < a/2$$

Hence the required probability is

$$P(|x|<a/2) = P\left(-\frac{a}{2}<X<\frac{a}{2}\right) = \int_{-a/2}^{a/2} dF(x)$$

$$= \frac{1}{2a}\int_{-a/2}^{a/2} 1.dx = \frac{1}{2}$$

Example 6·23. *Let $X_1, X_2, ..., X_n$ be a sequence of mutually independent random variables with common distribution. Suppose X_k assumes only positive integral values and $E(X_k) = a$, exists; $k = 1, 2, ..., n$. Let $S_n = X_1 + X_2 + \ldots + X_n$.*

(i) Show that $E\left(\frac{S_m}{S_n}\right) = \frac{m}{n}$, *for* $1 \le m \le n$

(ii) Show that $E\left(S_n^{-1}\right)$ *exists and*

$$E\left(\frac{S_m}{S_n}\right) = 1 + (m-n)\, a\, E(S_n^{-1}),\ \textit{for}\ 1 \le n \le m$$

(iii) Verify and use the inequality $x + x^{-1} \ge 2, (x > 0)$ *to show that*

$$E\left(\frac{S_m}{S_n}\right) \ge \frac{m}{n}\ \textit{for}\ m, n \ge 1$$

[Delhi Univ. M.Sc. (Stat.), 1988]

Solution. *(i)* We have

$$E\left[\frac{X_1+X_2+\ldots+X_n}{X_1+X_2+\ldots+X_n}\right] = E(1) = 1$$

$$\Rightarrow \quad E\left[\frac{X_1+X_2+\ldots+X_n}{S_n}\right] = 1$$

$$\Rightarrow \quad E\left(\frac{X_1}{S_n}\right)+E\left(\frac{X_2}{S_n}\right)+\ldots+E\left(\frac{X_n}{S_n}\right) = 1$$

Since X_i's, $(i = 1, 2, ..., n)$ are identically distributed random variables, (X_i/S_n), $(i = 1, 2, ..., n)$ are also identically distributed random variables.

$$\therefore \quad nE\left(\frac{X_i}{S_n}\right) = 1$$

$$\Rightarrow \quad E\left(\frac{X_i}{S_n}\right) = \frac{1}{n}\ ; \quad i = 1, 2, ..., n \qquad ...(*)$$

Now

$$E\left(\frac{S_m}{S_n}\right) = E\left(\frac{X_1+X_2+\ldots+X_m}{S_n}\right) = E\left[\frac{X_1}{S_n}+\frac{X_2}{S_n}+\ldots+\frac{X_m}{S_n}\right]$$

$$= E\left(\frac{X_1}{S_n}\right) + E\left(\frac{X_2}{S_n}\right) + \dots + E\left(\frac{X_m}{S_n}\right)$$

$$= \frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n} \text{ [(}m\text{ times)]} \qquad [Using\ (*)]$$

$$= \frac{m}{n}, (m < n)$$

(ii) Since X_i's assume only positive integral values, we have

$$n \le X_1 + X_2 + \dots + X_n < \infty$$

$$\Rightarrow \quad \frac{1}{n} \ge \frac{1}{S_n} > 0 \quad \Rightarrow \quad 0 < S_n^{-1} \le \frac{1}{n}$$

Since S_n^{-1} lies between two finite quantities 0 and $\frac{1}{n}$, we get

$$0 < E(S_n^{-1}) \le \frac{1}{n}$$

Hence $E(S_n^{-1})$ exists.

$$E\left(\frac{S_m}{S_n}\right) = E\left[\frac{X_1 + X_2 + \dots + X_n + X_{n+1} + \dots + X_m}{S_n}\right], \ m \ge n$$

$$= E\left[1 + \frac{X_{n+1}}{S_n} + \dots + \frac{X_m}{S_n}\right]$$

$$= 1 + E\left(\frac{X_{n+1}}{S_n}\right) + \dots + E\left(\frac{X_m}{S_n}\right)$$

Since $X_{n+1}, X_{n+2}, \dots, X_m$ are independent of $S_n = X_1 + X_2 + \dots + X_n$, they are independent of S_n^{-1} also.

$$\therefore \quad E\left(\frac{S_m}{S_n}\right) = 1 + E(X_{n+1}) E(S_n^{-1}) + \dots + E(X_m) E(S_n^{-1})$$

$$= 1 + [E(X_{n+1}) + \dots + E(X_m)] E(S_n^{-1})$$

$$= 1 + (m - n)\, a\, E(S_n^{-1}), \ 1 \le n \le m$$

$$[\because E(X_i) = a \ \forall \ i]$$

(iii) Verification of $x + \frac{1}{2} \ge 2, \ (x > 0)$.

$$x + \frac{1}{2} \ge 2$$

$$\Rightarrow \quad x^2 + 1 \ge 2x \quad \text{(multiplication valid only if } x > 0)$$

$$\Rightarrow \quad (x - 1)^2 \ge 0$$

which is always true for $x > 0$.

If $1 \le m \le n$, result follows from (*).

If $1 \le n \le m$, then using (**), we have to prove that

$$1 + (m-n)\,a\,E\,(S_n^{-1}) \geq \frac{m}{n}$$

$$\Rightarrow \quad (m-n)a\,E\,(S_n^{-1}) \geq \frac{m-n}{n}$$

$$\Rightarrow \quad E\,(S_n^{-1}) \geq \frac{1}{an} \quad \text{...(***)}$$

In (**), taking $x = \dfrac{S_n}{an} > 0$, we get

$$E\,(x) + E\,(x^{-1}) \geq 2$$

$$\Rightarrow \quad E\left[\frac{S_n}{an}\right] + E\left[\frac{S_n}{an}\right]^{-1} \geq 2$$

$$\Rightarrow \quad \frac{1}{an}.E\,(S_n) + an\,E\,(S_n^{-1}) \geq 2$$

$$\Rightarrow \quad \frac{1}{an}.an + an\,E\,(S_n^{-1}) \geq 2$$

$$\Rightarrow \quad an\,E\,(S_n^{-1}) \geq 1$$

$$\Rightarrow \quad E\,(S_n^{-1}) \geq \frac{1}{an},$$

which was to be proved in (***).

Example 6·24. *Let X be a r.v. for which β_1 and β_2 exist. Then for any real k, prove that :*

$$\beta_2 \geq \beta_1 - (2k + k^2) \quad \text{...(*)}$$

Deduce that (i) $\beta_2 \geq \beta_1$, (ii) $\beta_2 \geq 1$. When is $\beta_2 = 1$?

Solution. Without any loss of generality we can take $E\,(X) = 0$. [If $E\,(X) \neq 0$, then we may start with the random variable $Y = X - E\,(X)$ so that $E\,(Y) = 0$.]

Consider the real valued function of the real variable t defined by :

$$Z\,(t) = E\,[X^2 + tX + k\,\mu_2]^2 \geq 0 \quad \forall \quad t,$$

where $\mu_r = EX^r$, ...(*i*)

is the rth moment of X about mean.

$$\therefore \quad Z\,(t) = E\,[X^4 + t^2X^2 + k^2\mu_2^2 + 2tX^3 + 2k\,\mu_2 X^2 + 2k\,\mu_2\,t\,X]$$

$$= \mu_4 + t^2\mu_2 + k^2\mu_2^2 + 2t\,\mu_3 + 2k\,\mu_2^2$$

[Using (*i*) and $E\,(X) = 0$]

$$= t^2\mu_2 + 2t\,\mu_3 + \mu_4 + k^2\mu_2^2 + 2k\,\mu_2^2 \geq 0 \quad \text{for all } t. \quad \text{...(ii)}$$

Since $Z\,(t)$ is a quadratic form in t, $Z\,(t) \geq 0$ for all t iff its discriminant is ≤ 0, *i.e.*,

$$\text{iff} \quad (2\,\mu_3)^2 - 4\,\mu_2\,[\,\mu_4 + k^2\mu_2^2 + 2k\,\mu_2^2\,] \leq 0$$

$$\Rightarrow \quad \frac{\mu_3^2}{\mu_2^3} - \left[\frac{\mu_4}{\mu_2^2} + k^2 + 2k\right] \leq 0 \quad \text{[Dividing by } 4\,\mu_2^3 > 0]$$

$$\Rightarrow \qquad \beta_1 - \beta_2 - (2k + k^2) \le 0$$
$$\Rightarrow \qquad \beta_2 \ge \beta_1 - (2k + k^2)$$

Deductions. *(i)* Taking $k = 0$ in (*) we get $\beta_2 \ge \beta_1$ ...(**)

(ii) Taking $k = -1$ in (*) we get : $\beta_2 \ge \beta_1 + 1$ ...(***)

a result, which is established differently in Example 6·26.

(iii) Since $\beta_1 = \mu_3^2/\mu_2^3$ is always non-negative, we get from (***) :

$$\beta_2 \ge 1 \qquad ...(****)$$

Remark. The sign of equality holds in (****), *i.e.*, $\beta_2 = 1$ iff

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 1 \quad \Rightarrow \quad \mu_4 = \mu_2^2$$

$$\Rightarrow \qquad E[X - E(X)]^4 = [E(X - E(X))]^2$$
$$\Rightarrow \qquad E(Y^2) - [E(Y)]^2 = 0, \qquad (Y = [X - E(X)]^2)$$
$$\Rightarrow \qquad \text{Var}(Y) = 0$$
$$\Rightarrow \qquad P[Y = E(Y)] = 1 \qquad [\textit{See Example } 6{\cdot}9]$$
$$\Rightarrow \qquad P[(X - \mu)^2 = E(X - \mu)^2] = 1$$
$$\Rightarrow \qquad P[(X - \mu)^2 = \sigma^2] = 1$$
$$\Rightarrow \qquad P[(X - \mu) = \pm\sigma] = 1$$
$$\Rightarrow \qquad P[X = \mu \pm \sigma] = 1$$

Thus X takes only two values $\mu + \sigma$ and $\mu - \sigma$ with respective probabilities p and q, (say).

$$\therefore \qquad E(X) = p(\mu + \sigma) + q(\mu - \sigma) = \mu$$
$$\Rightarrow \qquad p + q = 1 \quad \text{and} \quad (p - q)\sigma = 0$$

But since $\sigma \ne 0$, (since in this case β_2 is defined) we have :

$$p + q = 1 \quad \text{and} \quad p - q = 0 \quad \Rightarrow \quad p = q = \tfrac{1}{2}.$$

Hence $\beta_2 = 1$ iff the r.v. X assumes only two values, each with equal probability ½.

Example 6·25. *Let X and Y be two variates having finite means.*

Prove or disprove :

(a) $E[\text{Min}(X, Y)] \le \text{Min}[E(X), E(Y)]$

(b) $E[\text{Max}(X, Y)] \ge \text{Max}[E(X), E(Y)]$

(c) $E[\text{Min}(X, Y) + \text{Max}(X, Y)] = E(X) + E(Y)$

[Delhi Univ. B.A. (Stat. Hons.), Spl. Course, 1989]

Solution. We know that

$$\text{Min}(X, Y) = \frac{1}{2}(X + Y) - |X - Y| \qquad ...(i)$$

and

$$\text{Max}(X, Y) = \frac{1}{2}(X + Y) + |X - Y| \qquad ...(ii)$$

(a) $\therefore \quad E[\text{Min}(X, Y)] = \frac{1}{2} E(X+Y) - E|X-Y| \qquad ...(iii)$

We have : $\quad |E(X-Y)| \leq E|X-Y|$

$\Rightarrow \quad -|E(X-Y)| \geq -E|X-Y|$

$\Rightarrow \quad E|X-Y| \leq -|E(X-Y)| = -|E(X)-E(Y)| \qquad ...(*)$

Substituting in *(iii)* we get :

$$E[\text{Min}(X, Y)] \leq \frac{1}{2} E(X+Y) - E|X-Y|$$

$$\leq \frac{1}{2}[E(X) + E(Y)] - |E(X) - E(Y)| \qquad ...[\text{From } (*)]$$

$\Rightarrow \quad E[\text{Min}(X, Y)] \leq \text{Min}[E(X), E(Y)]$

(b) Similarly from *(ii)* we get :

$$E[\text{Max}(X, Y)] = \frac{1}{2} E(X+Y) + E|X+Y|$$

$$\geq \frac{1}{2} E(X+Y) + |E(X+Y)|$$

$$(\because |E(X+Y)| \leq E|X+Y|)$$

$$= \frac{1}{2}[E(X) + E(Y)] + |E(X) + E(Y)|$$

$$= \text{Max}[E(X), E(Y)]$$

i.e., $\quad E[\text{Max}(X, Y)] \geq \text{Max}[E(X), E(Y)]$

(c) $\quad [\text{Min}(X, Y) + \text{Max}(X, Y)] = [X+Y]$

$\Rightarrow \quad E[\text{Min}(X, Y) + \text{Max}(X, Y)] = E(X+Y)$

$$= E(X) + E(Y),$$

as required.

Hence all the results in *(a)*, *(b)* and *(c)* are true.

Example 6·26. *Use the relation* $E(AX^a + BX^b + CX^c)^2 \geq 0$, *X being a random variable with* $E(X) = 0$, *E denoting the mathematical expectation, to show that*

$$\begin{vmatrix} \mu_{2a} & \mu_{a+b} & \mu_{a+c} \\ \mu_{a+b} & \mu_{2b} & \mu_{b+c} \\ \mu_{a+c} & \mu_{b+c} & \mu_{2c} \end{vmatrix} \geq 0, \qquad ...(*)$$

μ_n *denoting the nth moment about mean.*

Hence or otherwise show that Pearson Beta- coefficients satisfy the inequality

$$\beta_2 - \beta_1 - 1 \geq 0.$$

Also deduce that $\beta_2 \geq 1$.

Solution. Since $E(X) = 0$, we get

$$E(X^r) = \mu_r \qquad ...(**)$$

We are given that

$$E(AX^a + BX^b + CX^c)^2 \geq 0$$

$\Rightarrow \quad E[A^2X^{2a} + B^2X^{2b} + C^2X^{2c} + 2ABX^{a+b} + 2ACX^{a+c} + 2BCX^{b+c}] \geq 0$

$\Rightarrow \quad A^2\mu_{2a} + B^2\mu_{2b} + C^2\mu_{2c} + 2AB\mu_{a+b} + 2AC\mu_{a+c} + 2BC\mu_{b+c} \geq 0$

[From (**)] ...(***)

for all values of A, B, C.

We know from matrix theory that the conditions for the quadratic form

$$a'x^2 + b'y^2 + c'z^2 + 2f'yz + 2g'zx + 2h'xy,$$

to be non-negative for all values of x, y and z are

$$a' \geq 0, \quad (ii) \begin{vmatrix} a' & h' \\ h' & b' \end{vmatrix} \geq 0, \text{ and } \quad (iii) \begin{vmatrix} a' & h' & g' \\ h' & b' & f' \\ g' & f' & c' \end{vmatrix} \geq 0$$

Comparing with (***), we have

$$a' = \mu_{2a}, b' = \mu_{2b}, c' = \mu_{2c}, f' = \mu_{b+c}, g' = \mu_{a+c}, h' = \mu_{a+b}$$

Substituting these values in condition *(iii)* above, we get the required result.

Taking $a = 0, b = 1$ and $c = 2$ in (*) and noting that $\mu_0 = 1$ and $\mu_1 = 0$, we get

$$\begin{vmatrix} 1 & 0 & \mu_2 \\ 0 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix} \geq 0$$

$$\Rightarrow \qquad \mu_2 \mu_4 - \mu_3^2 + \mu_2(-\mu_2^2) \geq 0$$

Dividing throughout by μ_2^3 (assuming that μ_2 is finite, for otherwise β_2 will become infinite), we get

$$\frac{\mu_4}{\mu_2^2} - \frac{\mu_3^2}{\mu_2^3} - 1 \geq 0$$

$$\Rightarrow \qquad \beta_2 - \beta_1 - 1 \geq 0$$

$$\Rightarrow \qquad \beta_2 \geq \beta_1 + 1$$

Further since $\beta_1 \geq 0$, we get $\beta_2 \geq 1$.

Example 6·27. *Let X be a non-negative random variable with distribution function F. Show that*

$$E(X) = \int_0^\infty [1 - F(x)]\, dx. \qquad ...(i)$$

Conjecture a corresponding expression for $E(X^2)$.

[Delhi Univ. M.Sc.(Stat). 1988]

Solution. Since $X \geq 0$, we have :

$$\text{R.H.S.} = \int_0^\infty [1 - P(X \leq x)]\, dx = \int_0^\infty \left[1 - \int_0^x f(u)\, du\right] dx,$$

where $f(\cdot)$ is the p.d.f. of r.v. X.

$$\therefore \qquad \text{R.H.S.} = \int_0^\infty \left[\int_x^\infty f(u)\, du\right] dx; \qquad ...(ii)$$

From the integral in bracket *(ii)*, we have, $u \geq x$ and since x ranges from 0 to ∞, u also range from 0 to ∞. Further $u \geq x \Rightarrow x \leq u$ and since x is non-negative, we have $0 \le x \le u$. [See Region R_1 in Remark 2 below]. Hence changing the order of integration in *(ii)*, [by Fubini's theorem for non-negative functions], we get

$$\text{R.H.S.} = \int_0^\infty \left[\int_0^u 1 \cdot dx \right] f(u)\, du = \int_0^\infty u \cdot f(u)\, du$$

$$= E(X) \qquad \qquad [\textit{Since } f(\cdot) \textit{ is p.d.f. of } X]$$

Conjecture for $E(X^2)$. Consider the integral :

$$\int_0^\infty 2x[1 - F(x)]\, dx = \int_0^\infty 2x \left(\int_x^\infty f(u)\, du \right) dx$$

$$= \int_0^\infty \left(\int_0^u 2x\, dx \right) f(u)\, du,$$

(By Fubini's theorem for non-negative functions).

$$= \int_0^\infty u^2 \cdot f(u) \cdot du = E(X^2)$$

Remarks. 1. If X is a non-negative r.v. then

$$\text{Var}\, X = EX^2 - [E(X)]^2 = \int_0^\infty 2x[1 - F(x)]\, dx - \mu_x^2 \qquad \text{...(iii)}$$

2. If we do not restrict ourselves to non-negative random variables only, we have the following more generalised result.

If F denotes the distribution function of the random variable X then :

$$E(X) = \int_0^\infty [1 - F(x)]\, dx - \int_{-\infty}^0 F(x)\, dx, \qquad \text{...(iv)}$$

provided the integrals exist finitely.

Proof of (iv). The first integral has already been evaluated in the above example, *i.e.*,

$$\int_0^\infty [1 - F(x)]\, dx = \int_0^\infty u \cdot f(u) \cdot du \qquad \text{...(v)}$$

Consider :

$$\int_{-\infty}^0 F(x)\, dx = \int_{-\infty}^0 P(X \le x)\, dx = \int_{-\infty}^0 \left(\int_{-\infty}^x f(u)\, du \right) dx$$

$$= \int_{-\infty}^0 \left(\int_u^0 1 \cdot dx \right) f(u)\, du$$

[Changing the order of integration in the Region R_2 where $u \le x$].

$$= \int_{-\infty}^{0} u \cdot f(u)\, du \qquad \text{...(vi)}$$

Subtracting *(vi)* from *(v)*, we get :

$$\int_0^{\infty} [1 - F(x)]\, dx - \int_{-\infty}^{0} F(x)\, dx = \int_0^{\infty} u f(u)\, du + \int_{-\infty}^{0} u f(u)\, du$$

$$= \int_{-\infty}^{\infty} u f(u)\, du$$

$$= E(X), \qquad (\text{Since } f(\cdot) \text{ is p.d.f. of } X)$$

as desired.

In this generalised case,

$$\text{Var}(X) = \int_0^{\infty} 2x\, [1 - F_X(x) + F_X(-x)]\, dx - [E(X)]^2$$

3. The corresponding analogue of the above result for discrete random variable is given in the next Example 6·28.

Example 6·28. *If the possible values of a variate X are 0, 1, 2, 3, ... then*

$$E(X) = \sum_{n=0}^{\infty} P(X > n)$$

[Delhi Univ. B.Sc. (Maths Hons.), 1987]

Solution. Let $P(X = n) = p_n,\ n = 0, 1, 2, 3, \ldots$ ...*(i)*

If $E(X)$ exists, then by definition :

$$E(X) = \sum_{n=0}^{\infty} n \,.\, P(X = n) = \sum_{n=1}^{\infty} n \,.\, p_n \qquad \text{...(ii)}$$

Consider :

$$\sum_{n=0}^{\infty} P(X > n) = P(X > 0) + P(X > 1) + P(X > 2) + \ldots$$

$$= P(X \geq 1) + P(X \geq 2) + P(X \geq 3) + \ldots$$

$$= (p_1 + p_2 + p_3 + p_4 + \ldots\ldots)$$
$$+ (p_2 + p_3 + p_4 + \ldots\ldots)$$
$$+ (p_3 + p_4 + p_5 + \ldots\ldots)$$
$$+ \ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$= p_1 + 2p_2 + 3p_3 + \ldots\ldots$$

$$= \sum_{n=1}^{\infty} n\, p_n$$

$$= E(X) \qquad [\text{From } (ii)]$$

As an illustration of this result, see Problem 24 in Exercise 6(a).

Aliter. R.H.S. $= \sum_{n=0}^{\infty} P(X > n) = \sum_{n=1}^{\infty} P(X \geq n)$

$$= \sum_{n=1}^{\infty} \left(\sum_{x=n}^{\infty} p(x) \right)$$

Since the series is convergent and $p(x) \geq 0 \ \forall \ x$, by Fubini's theorem, changing the order of summation we get :

$$\text{R.H.S.} = \sum_{x=1}^{\infty} \left(\sum_{n=1}^{x} p(x) \right) = \sum_{x=1}^{\infty} \left\{ p(x) \sum_{n=1}^{x} 1 \right\}$$

since $x \geq n \ \Rightarrow \ n \leq x$ and x assumes only positive integral values :

$$\therefore \qquad \text{R.H.S.} = \sum_{x=1}^{\infty} x\,p(x) = \sum_{x=0}^{\infty} x\,p(x) = E(X)$$

Example 6·29. *For any variates X and Y, show that*

$$\left\{ E(X+Y)^2 \right\}^{1/2} \leq \left\{ E(X^2) \right\}^{1/2} + \left\{ E(Y^2) \right\}^{1/2} \qquad ...(*)$$

Solution. Squaring both sides in (*), we have to prove

$$E(X+Y)^2 \leq \left[\left\{ E(X^2) \right\}^{1/2} + \left\{ E(Y^2) \right\}^{1/2} \right]^2$$

$$\Rightarrow \quad E(X^2) + E(Y^2) + 2E(XY) \leq E(X^2) + E(Y^2) + 2\sqrt{E(X^2)\,E(Y^2)}$$

$$\Rightarrow \qquad E(XY) \leq \sqrt{E(X^2)\,E(Y^2)}$$

$$\Rightarrow \qquad [E(XY)]^2 \leq E(X^2)\,.\,E(Y^2).$$

This is nothing but Cauchy-Schwartz inequality. [For proof see Theorem 6·11 page 6·13.]

Example 6·30. *Let X and Y be independent non-degenerate variates. Prove that*

$$Var(XY) = Var(X)\,.\,Var(Y)$$

iff $\qquad E(X) = 0, \ E(Y) = 0$

[Delhi Univ. B.Sc. (Maths Hons.), 1989]

Solution. We have

$$\text{Var}(XY) = E(XY)^2 - [E(XY)]^2 = E(X^2Y^2) - [E(XY)]^2$$

$$= E(X^2)\,E(Y^2) - [E(X)]^2\,[E(Y)]^2, \qquad ...(*)$$

since X and Y are independent.

$$\left.\begin{array}{ll} \text{If} & E(X) = 0 = E(Y) \\ \text{then} & \text{Var}(X) = E(X^2) \text{ and } \text{Var}(Y) = E(Y^2) \end{array}\right\} \qquad ...(**)$$

Substituting from (**) in (*), we get

$$\text{Var}(XY) = \text{Var}(X)\,.\,\text{Var}(Y),$$

as desired.

Only If. We have to prove that if

$$\text{Var}(XY) = \text{Var}(X)\,.\,\text{Var}(Y) \qquad ...(***)$$

then $\qquad E(X) = 0$ and $E(Y) = 0$.

Now (***) gives, [on using (*)]

$$E(X^2)E(Y^2) - [E(X)]^2.[E(Y)]^2 = \left\{ E(X^2) - [E(X)]^2 \right\} \times \left\{ E(Y^2) - [E(Y)]^2 \right\}$$

$$= E(X^2)E(Y^2) - E(X^2)[E(Y)]^2 - [E(X)]^2 E(Y^2) + [E(X)]^2[E(Y)]^2$$

$$\Rightarrow E(X^2).[E(Y)]^2 - [E(X)]^2.[E(Y)]^2 + [E(X)]^2.E(Y^2) - [E(X)]^2[E(Y)]^2 = 0$$

$$\Rightarrow [E(Y)]^2\{E(X^2) - [E(X)]^2\} + [E(X)]^2\{E(Y^2) - [E(Y)]^2\} = 0$$

$$\Rightarrow \quad [E(Y)]^2 \text{Var}(X) + [E(X)]^2 \text{Var}(Y) = 0 \qquad \ldots(****)$$

Since each of the quantities $[E(X)]^2$, $[E(Y)]^2$, Var (X) and Var (Y) is non-negative and since X and Y are given to be non-degenerate random variables such that Var $(X) > 0$ and Var $(Y) > 0$, (****) holds only if we have $E(X) = 0 = E(Y)$, as required.

EXERCISE 6(a)

1. *(a)* Define a random variable and its mathematical expectation.

(b) Show that the mathematical expectation of the sum of two random variables is the sum of their individual expectations and if two variables are independent, the mathematical expectation of their product is the product of their expectations.

Is the condition of independence necessary for the latter ? If not, what is the necessary condition ?

(c) If X is a random variable, prove that $|E(X)| \le E(|X|)$.

(d) If X and Y are two random variables such that $X \le Y$, prove that

$$E(X) \le E(Y).$$

(e) Prove that $E[(X-c)^2] = [\text{Var}(X)] + [E(X) - c]^2$, where c is a constant.

2. Prove that

(a) $E(aX + bY) = a\,E(X) + b\,E(Y)$.

where a and b are any constants.

(b) $E(a) = a$, a being a constant.

(c) $E[ag(X)] = aE[g(X)]$

(d) $E[g_1(X) + g_2(X) + \ldots + g_n(X)] = E[g_1(X)] + E[g_2(X)] + \ldots + E[g_n(X)]$

(e) $|E[g(X)]| \le E[|g(X)|]$.

(f) If $g(X) \ge 0$, everywhere then $E[g(X)] \ge 0$.

(g) If $g(X) \ge 0$ everywhere and $E[g(X)] = 0$, then $g(X) = 0$, *i.e.*, the random variable $g(X)$ has a one point distribution at $X = 0$.

3. Show that if X is non-negative random variable such that both $E(X)$ and $E(1/X)$ exist, then

$$E(1/X) \ge 1/E(X).$$

4. If X, Y are independent random variables with $E(X) = \alpha$, $E(X^2) = \beta$, and $E(Y^k) = a_k$; $k = 1, 2, 3, 4$, find $E(XY + Y^2)^2$.

5. *(a)* If X and Y are two independent random variables, show that

$$\text{Var}(aX + bY) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y).$$

(b) With usual notations, show that

$\text{Cov}(aX + bY, cX + dY) = ac \text{ Var}(X) + bd \text{ Var}(Y) + (ad + bc) \text{ Cov}(X, Y)$

(c) Show that

$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{n} b_j X_j\right) = \sum_{j=1}^{n}\sum_{i=1}^{n} a_i b_j \text{ Cov}(X_i, X_j)$$

6. (a) Define the indicator function $I_A(x)$ and show that $E(I_A(X)) = P(A)$.

(b) Prove that the probability function $P(X \in A)$ for set A and the distribution function $F_X(x)$, $(-\infty < x < \infty)$, can be regarded as expectations of some random variable.

Hint. Define the indicator functions :

$$I_A(x) = 1 \text{ if } x \in A \qquad\qquad I_y(x) = 1 \text{ if } x \le y$$
$$= 0 \text{ if } x \notin A \qquad\qquad = 0 \text{ if } x > y$$

Then we shall get :

$E(I_A(X)] = P(X \in A)$ and $E(I_y(X)) = P(X \le y) = F_X(y)$

7. (a) Let X be a continuous random variable with median m. Minimise $E|X - b|$, as an function of b.

Ans. $E|X - b|$ is minimum when $b = m =$ Median. This states that absolute sum of deviations of a given set of observations is minimum when taken about median. [See Example 5·19.]

(b) Let X be a random variable such that $E|X| < \infty$. Show that $E|X - C|$ is minimised if we choose C equal to the median of the distribution.

[Delhi Univ. B.Sc. (Maths Hons.), 1988]

8. If X and Y are symmetric, show that

$$E\left(\frac{X}{X+Y}\right) = \frac{1}{2}$$

Hint. $$1 = E\left[\frac{X+Y}{X+Y}\right] = E\left[\frac{X}{X+Y}\right] + E\left[\frac{Y}{X+Y}\right]$$

$$\Rightarrow \qquad 1 = 2E\left[\frac{X}{X+Y}\right] \qquad (\because X \text{ and } Y \text{ are symmetric.})$$

9. (a) If a r.v. X has a symmetric density about the point 'a' and if $E(X)$ exists, then

$$\text{Mean}(X) = \text{Median}(X) = a$$

Hint. Given $f(a - x) = f(a + x)$; $f(x)$ p.d.f. of X. prove that

$$E(X - a) = \int_{-\infty}^{\infty}(x - a)f(x)\,dx = \int_{-\infty}^{\infty}(x - a)f(x)\,dx + \int_{a}^{\infty}(x - a)f(x)\,dx = 0$$

(b) If X and Y are two random variables with finite variances, then show that

$$E^2(XY) \le E(X^2) \cdot E(Y^2) \qquad ...(*)$$

When does the equality sign hold in (*) ? **[Indian Civil Service, 1987]**

10. Let X be a non-negative arbitrary r.v. with distribution function F. show that

$$E(X) = \int_0^{\infty} [1 - F_X(x)]\, dx - \int_{-\infty}^{0} F_X(x)\, dx,$$

in the sense that, if either side exists, so does the other and the two are equal.

[Delhi Univ. B.Sc. (Maths Hons.), 1992]

11. Show that if Y and Z are independent random values of a variable X, the expected value of $(Y - Z)^2$ is twice the variance of the distribution of X.

[Allahabad Univ. B.Sc., 1989]

Hint. $E(Y) = E(Z) = E(X) = \mu$, (say) ; $\sigma_y^2 = \sigma_z^2 = \sigma_x^2 = \sigma^2$, (say) . ...(*)

$$E(Y - Z)^2 = E(Y^2) + E(Z^2) - 2E(Y)E(Z)$$

($\because$ Y, Z are independent)

$$= (\sigma_y^2 + \mu_y^2) + (\sigma_z^2 + \mu_z^2) - 2\mu^2$$

$$= 2\sigma^2 = 2\sigma_x^2$$ [On using (*)]

12. Given the following table :

x	-3	-2	-1	0	1	2	3
$p(x)$	0·05	0·10	0·30	0	0·30	0·15	0·10

Compute (*i*) $E(X)$, (*ii*) $E(2X \pm 3)$, (*iii*) $E(4X + 5)$, (*iv*) $E(X^2)$

(*v*) $V(X)$, and (*vi*) $V(2X \pm 3)$.

13. (*a*) A and B throw with one die for a stake of Rs. 44 which is to be won by the playr who first throws a 6. If A has the first throw, what are their respective expectations?

Ans. Rs. 24, Rs. 20.

(*b*) A contractor has to choose between two jobs. The first promises a profit of Rs. 1,20,000 with a probability of ¾ or a loss of Rs. 30,000 due to delays with a probability of ¼ ; the second promises a profit of Rs. 1,80,000 with a probability of ½ or a loss of Rs. 45,000 with a probability of ½. Which job should the contractor choose so as to maximise his expected profit?

(*c*) A random variable X can assume any positive integral value n with a probability porportional to $1/3^n$. Find the expectation of X.

[Delhi Univ. B.Sc., Oct. 1987]

14. Three tickets are chosen at random without replacement from 100 tickets numbered 1, 2, ..., 100. Find the mathematical expectation of the sum of the numbers on the tickets drawn.

15. (*a*) Three urns contain respectavily 3 green and 2 white balls, 5 green and 6 white balls and 2 green and 4 white balls. One ball is drawn from each urn. Find the expected number of white balls drawn out.

Hint. Let us define the r.v.

$X_i = 1$, if the ball drawn from ith urn is white
$= 0$; otherwise

Then the number of white balls drawn is $S = X_1 + X_2 + X_3$.

$$E(S) = E(X_1) + E(X_2) + E(X_3) = 1 \times \frac{2}{5} + 1 \times \frac{6}{11} + 1 \times \frac{4}{6} = \frac{266}{165}.$$

(b) Urn A contains 5 cards numbered from 1 to 5 and urn B contains 4 cards numbered from 1 to 4. One card is drawn from each of these urns. Find the probability function of the number which appears on the cards drawn and its mathematical expectation.

Ans. 11/4.

16. *(a)* Thirteen cards are drawn from its pack simultaneously. If the values of aces are 1, face cards 10 and others according to denomination, find the expectation of the total score in all the 13 cards.

[Madurai Univ. B.Sc., Oct. 1990]

(b) Let X be a random variable with p.d.f. as given below :

x :	0	1	2	3
$p(x)$:	1/3	1/2	1/24	1/8

Find the expected value of $Y = (X - 1)^2$. **[Aligarh Univ. B.Sc. (Hons.), 1992]**

17. A player tosses 3 fair coins. He wins Rs. 8, if three heads occur; Rs. 3, if 2 heads occur and Re. 1, if one head occurs. If the game is to be fair, how much should he lose, if no heads occur? **[Punjab Univ. M.A. (Econ.), 1987]**

Hint. X : Player's prize in Rs.

x	8	3	1	a
No. of heads	3	2	1	0
$p(x)$	⅛	⅜	⅜	⅛

$E(X) = \Sigma x\, p(x) = \frac{1}{8}(8 + 9 + 3 + a)$

For the game to be fair, we have :

$E(X) = 0 \Rightarrow 20 + a = 0 \Rightarrow a = -20$

Hence the player loses Rs. 20, if no heads come up.

18. *(a)* A coin is tossed until a tail appears. What is the expectation of the number of tosses ?

Ans. 2.

(b) Find the expectation of *(i)* the sum, and *(ii)* the product, of number of points on n dice when thrown.

Ans. *(i)* $7n/2$, *(ii)* $(7/2)^n$

19. *(a)* Two cards are drawn at random from ten cards numbered 1 to 10. Find the expectation of the sum of points on two cards.

(b) An urn contains n cards marked from 1 to n. Two cards are drawn at a time. Find the mathematical expectation of the product of the numbers on the cards.

[Mysore Univ. B.Sc., 1991]

(c) In a lottery m tickets are drawn out of n tickets numbered from 1 to n. What is the expectation of the sum of the squares of numbers drawn ?

(d) A bag contains n white and 2 black balls. Balls are drawn one by one without replacement until a black is drawn. If 0, 1, 2, 3, ... white balls are drawn before the first black, a man is to receive $0^2, 1^2, 2^2, 3^2, \ldots$ rupees respectively. Find his expectation. **[Rajasthan Univ. B.Sc., 1992]**

(e) Find the expectation and variance of the number of successes in a series of independent trials, the probability of success in the ith trial being p_i ($i = 1, 2, \ldots, n$). **[Nagarjuna Univ. B.Sc., 1991]**

20. Balls are taken one by one out of an urn containing w white and b black balls until the first white ball is drawn. Prove that the expectation of the number of black balls preceding the first white ball is $b/(w+1)$.

[Allahabad Univ. B.Sc. (Hons.), 1992]

21. *(a)* X and Y are independent variables with means 10 and 20, and variances 2 and 3 respectively. Find the variance of $3X + 4Y$.

Ans. 66.

(b) Obtain the variance of $Y = 2X_1 + 3X_2 + 4X_3$ where X_1, X_2 and X_3 are three random variables with means given by 3, 4, 5 respectively, variances by 10, 20, 30 respectively, and co-variances by $\sigma_{X_1X_2} = 0$, $\sigma_{X_2X_3} = 0$, $\sigma_{X_1X_3} = 5$, where $\sigma_{X_rX_s}$ stands for the co-variance of X_r and X_s.

22. *(a)* Suppose that X is a random variable for which $E(X) = 10$ and $\text{Var}(X) = 25$. Find the positive values of a and b such that $Y = aX - b$, has expectation 0 and variance 1.

Ans. $a = 1/5,\ b = 2$

(b) Let X_1 and X_2 be two stochastic random variables having variances k and 2 respectively. If the variance of $Y = 3X_2 - X_1$ is 25, find k.

(Poona Univ. B.Sc., 1990)

Ans. $k = 7$.

23. A bag contains $2n$ counters, of which half are marked with odd numbers and half with even numbers, the sum of all the numbers being S. A man is to draw two counters. If the sum of the numbers drawn is odd, he is to receive that number of rupees, if even he is to pay that number of rupees. Show that his expectation is $S/[n(2n-1)]$ rupees. **(I.F.S., 1989)**

24. A jar has n chips numbred 1, 2, ..., n. A person draws a chip, returns it, draws another, returns it, and so on, until a chip is drawn that has been drawn before. Let X be the number of drawings. Find $E(X)$.

[Delhi Univ. B.A. (Stat. Hons.), Spl. Course, 1986]

Hint. Obviously $P(X > 1) = 1$, because we must have at least two draws to get the chip which has been drawn before.

$$P(X > r) = P[\text{Distinct number on } i\text{th draw; } i = 1, 2, \ldots, r]$$

$$= \frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \dots \times \frac{n-(r-1)}{n}$$

$$P(X > r) \Rightarrow \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \dots \dots \left(1 - \frac{r-1}{n}\right); r = 1, 2, 3, \dots \qquad \dots(*)$$

Hence, using the result in Example 6·28

$$E(X) = \sum_{r=0}^{\infty} P(X > r)$$

$$= P(X > 0) + P(X > 1) + P(X > 2) + \dots$$

$$= 1 + 1 + \left(1 - \frac{1}{n}\right) + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) +$$

$$\dots + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{r-1}{n}\right) + \dots \qquad \text{[Using (*)]}$$

25. A coin is tossed four times. Let X denote the number of times a head is followed immediately by a tail. Find the distribution, mean and variance of X.

Hint. $S = \{H, T\} \times \{H, t\} \times \{H, T\} \times \{H, T\}$

$= \{HHHH, HHHT, HHTH, HTHH, HTHT, \dots\dots, TTTT\}$

X: 0, 1, 1, 1, 2, ..., 0

x	0	1	2
$p(x)$	5/16	10/16	1/16

$E(X) = 3/4$, $E(X^2) = 7/8$

$\text{Var } X = 7/8 - 9/16 = 5/16$.

26. An urn contains balls numbered 1, 2, 3. First a ball is drawn from the urn and then a fair coin is tossed the number of times as the number shown on the drawn ball. Find the expected number of heads.

[Delhi Univ. B.Sc. (Maths Hons.), 1984]

Hint. B_j : Event of drawing the ball numbered j.

$P(B_j) = 1/3$; $j = 1, 2, 3$.

X : No. of heads shown. X is a r.v. taking the values 0, 1, 2, and 3.

$$P(X = x) = \sum_{j=1}^{3} P(B_j) \cdot P(X = x \mid B_j) = \frac{1}{3} \sum_{j=1}^{3} P(X = x \mid B_j)$$

$$\therefore P(X = 0) = \frac{1}{3}\left[P(X = 0 \mid B_1) + P(X = 0 \mid B_2) + P(X = 0 \mid B_3)\right]$$

$$= \frac{1}{3}\left[\frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right] = \frac{7}{24}$$

$$P(X = 1) = \frac{1}{3}\left[P(X = 1 \mid B_1) + P(X = 1 \mid B_2) + P(X = 1 \mid B_3)\right]$$

$$= \frac{1}{3}\left[\frac{1}{2} + \frac{2}{4} + \frac{3}{8}\right]$$

e.g., $P(X = 0 \mid B_2) = P$ [No head when two coins are tossed] $= 2/4$

$P(X = 1 \mid B_3) = P$ [1 head when three coins are tossed] $= 3/8$

Similarly $P(X = 2) = \frac{1}{3}\left(0 + \frac{1}{4} + \frac{3}{8}\right) = \frac{5}{24}$

$$P(X=3)=\frac{1}{3}\left(0+0+\frac{1}{8}\right)=\frac{1}{24}$$

$$\therefore \quad E(X)=\sum_{x=0}^{3} x\,P(X=x)=\frac{11}{24}+\frac{10}{24}+\frac{3}{24}=1$$

27. An urn contains pN white and qN black balls, the total number of balls being N, $p+q=1$. Balls are drawn one by one (without being returned to the urn) until a certain number n of balls is reached.

Let $X_i = 1$, if the ith ball drawn is white.
$= 0$, if the ith ball drawn is black.

(*i*) Show that $E(X_i)=p$, $\text{Var}(X_i)=pq$.

(*ii*) Show that the co-variance between X_j and X_k is $-\dfrac{pq}{n-1}$, $(j\neq k)$

(*iii*) From (*i*) and (*ii*), obtain the variance of $S_n = X_1 + X_2 + \ldots + X_n$.

28. Two similar decks of n distinct cards each are put into random order and are matched against each other. Prove that the probability of having exactly r matches is given by

$$\frac{1}{r\,!}\sum_{k=0}^{n-r}\frac{(-1)^k}{k\,!},\quad r=0,1,2,\ldots n$$

Prove further that the expected number of matches and its variance are equal and are independent of n.

29. (*a*) If X and Y are two independent random variables, such that $E(X)=\lambda_1$, $V(X)=\sigma_1^2$ and $E(Y)=\lambda_2$, $V(Y)=\sigma_2^2$, then prove that

$$V(XY)=\sigma_1^2\sigma_2^2+\lambda_1^2\sigma_2^2+\lambda_2^2\sigma_1^2$$ **[Gorakhpur Univ. B.Sc., 1992]**

(*b*) If X and Y are two independent random variables, show that

$$\frac{V(XY)}{[E(X)]^2[E(Y)]^2}=C_X^2C_Y^2+C_X^2+C_Y^2$$

where $$C_X=\frac{\sqrt{V(X)}}{E(X)},\quad C_Y=\frac{\sqrt{V(Y)}}{E(Y)}$$

are the so-called coefficients of variation of X and Y? **[Patna Univ. B.Sc., 1991]**

30. A point P is taken at random in a line AB of length $2a$, all positions of the point being equally likely. Show that the expected value of the area of the rectangle $AP.\,PB$ is $2a^2/3$ and the probability of the area exceeding $1/2a^2$ is $1/\sqrt{2}$. **[Delhi Univ. B.Sc. (Maths Hons.), 1986]**

31. If X is a random variable with $E(X)=\mu$ satisfying $P(X\leq 0)=0$, show that $P(X>2\mu)\leq 1/2$. **[Delhi Univ. B.Sc. (Maths Hons.), 1992]**

OBJECTIVE TYPE QUESTIONS

1. Fill in the blanks :

(*i*) Expected value of a random variable X exists if ....

(*ii*) If $E(X^r)$ exists then $E(X^s)$ also exists for ....

(*iii*) When X is a random variable, expectation of $(X-\text{constant})^2$ is minimum when the constant is ....

(iv) $E\,|X - A|$ is minimum when $A =$ ....

(v) Var $(c) =$ ..., where c is a constant

(vi) Var $(X + c) =$ ..., where c is a constant

(vii) Var $(aX + b) =$ ..., where a and b are constants.

(viii) If X is a r.v. with mean μ and variance σ^2 then

$$E\left(\frac{X-\mu}{\sigma}\right) = \ldots, \quad \text{Var}\left(\frac{X-\mu}{\sigma}\right) = \ldots$$

(ix) $[E(XY)]^2 \ldots E(X^2)\,.\,E(Y^2)$.

(x) $V(aX \pm bY) = \ldots$

where a and b are constants.

II. Mark the correct answer in the following :

(i) For two random variables X and Y, the relation

$$E(XY) = E(X)\,E(Y)$$

holds good

(a) if X and Y are statistically independent,

(b) for all X and Y,

(c) if X and Y are identical.

(ii) Var $(2X \pm 3)$ is

(a) 5 *(b)* 13 *(c)* 4, if Var $X = 1$.

(iii) $E(X - k)^2$ is minimum when

(a) $k < E(X)$, *(b)* $k > E(X)$, *(c)* $k = E(X)$.

III. Comment on the following :

If X and Y are mutually independent variables, then

(i) $E(XY + Y + 1) - E(X + 1)\,E(Y) = 0$

(ii) X and Y are independent if and only if

$$\text{Cov}(X, Y) = 0$$

(iii) For every univariable distribution :

(a) $V(cX) = c^2 V(X)$ *(b)* $E(c/X) = c/E(X)$

(iv) Expected value of a r.v. always exists.

IV. Mark true or false with reasons for your answers :

(a) Cov $(X, Y) = 0 \Rightarrow X$ and Y are independent.

(b) If Var $(X) >$ Var (Y), then $X + Y$ and $X - Y$ are dependent.

(c) If Var $(X) =$ Var (Y) and if $2X + Y$ and $X - Y$ are independent, then X and Y are dependent.

(d) If Cov $(aX + bY, bX + aY) \neq ab$ Var $(X + Y)$, then X and Y are dependent.

6·8. Moments of Bivariate Probability Distributions. The mathematical expectation of a function $g(x, y)$ of two-dimensional random variable (X, Y) with p.d.f. $f(x, y)$ is given by

$$E[g(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy \qquad ...(6\cdot43)$$

(If X and Y are continuous variables)

$$= \sum_i \sum_j x_i y_j P(X = x_i \cap Y = y_j), \qquad ...(6\cdot43\ a)$$

(If X and Y are discrete variables)

provided the expectation exists.

In particular, the rth and sth product moment about origin of the random variables X and Y respectively is defined as

$$\mu_{rs}' = E(X^r Y^s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^r y^s f(x, y)\, dx\, dy$$

or $$\mu_{rs}' = \sum_i \sum_j x_i^r y_j^s P(X = x_i \cap Y = y_j) \qquad ...(6\cdot44)$$

The joint rth central moment of X and sth central moment of Y is given by

$$\mu_{rs} = E\left[\{X - E(X)\}^r \{Y - E(Y)\}^s\right]$$

$$= E\left[(X - \mu_X)^r (Y - \mu_Y)^s\right], \quad [E(X) = \mu_X, E(Y) = \mu_Y] \qquad ...(6\cdot45)$$

In particular

$$\mu_{00}' = 1 = \mu_{00}, \quad \mu_{10} = 0 = \mu_{01}$$
$$\mu_{10}' = E(X), \quad \mu_{01}' = E(Y)$$
$$\mu_{20} = \sigma_X^2, \quad \mu_{02} = \sigma_Y^2 \text{ and } \mu_{11} = \text{Cov}(X, Y).$$

6·9. Conditional Expectation and Conditional Variance.

Discrete Case. The conditional expectation or mean value of a continuous function $g(X, Y)$ given that $Y = y_j$, is defined by

$$E\{g(X, Y) \mid Y = y_j\} = \sum_{i=1}^{\infty} g(x_i, y_j)\, P(X = x_i \mid Y = y_j)$$

$$= \frac{\sum_{i=1}^{\infty} g(x_i, y_j)\, P(X = x_i \cap Y = y_j)}{P(Y = y_j)} \qquad ...(6\cdot46)$$

i.e., $E[g(X, Y) \mid Y = y_j]$ is nothing but the expectation of the function $g(X, y_j)$ of X w.r.t. the conditional distribution of X when $Y = y_j$.

In particular, the conditional expectation of a discrete random variable X given $Y = y_j$ is

$$E(X \mid Y = y_j) = \sum_{i=1}^{\infty} x_i\, P(X = x_i \mid Y = y_j) \qquad ...(6\cdot47)$$

The conditional variance of X given $Y = y_j$ is likewise given by

$$V(X \mid Y = y_j) = E[\{X - E(X \mid Y = y_j)\}^2 \mid Y = y_j] \qquad ...(6\cdot47a)$$

The conditional expectation of $g(X, Y)$ and the conditional variance of Y given $X = x_i$ may also be defined in an exactly similar manner.

Continuous Case. The conditional expectation of $g(X, Y)$ on the hypothesis $Y = y$ is defined by

$$E\{g(X, Y) \mid Y = y\} = \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x \mid y)\, dx$$

$$= \frac{\int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx}{f_Y(y)} \qquad \ldots(6{\cdot}48)$$

In particular, the conditional mean of X given $Y = y$ is defined by

$$E(X \mid Y = y) = \frac{\int_{-\infty}^{\infty} x f(x, y)\, dx}{f_Y(y)}$$

Similarly, we define

$$E(Y \mid X = x) = \frac{\int_{-\infty}^{\infty} y f(x, y)\, dy}{f_X(x)} \qquad \ldots(6{\cdot}48\ a)$$

The conditional variance of X may be defined as

$$V(X \mid Y = y) = E\left[\{X - E(X \mid Y = y)\}^2 \mid Y = y\right]$$

Similarly, we define

$$V(Y \mid X = x) = E\left[\{Y - E(Y \mid X = x)\}^2 \mid X = x\right] \qquad \ldots(6{\cdot}49)$$

Theorem 6·13. *The expected value of X is equal to the expectation of the conditional expectation of X given Y. Symbolically,*

$$E(X) = E[E(X \mid Y)] \qquad \ldots(6{\cdot}50)$$

[Calicut Univ. B.Sc. (Main Stat.), 1980]

Proof. $E[E(X \mid Y)] = E[\sum_i x_i\ P(X = x_i \mid Y = y_j)]$

$$= E\left[\sum_i x_i \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}\right]$$

$$= \sum_j \left[\sum_i \left\{x_i \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}\right\}\right] P(Y = y_j)$$

$$= \sum_j \sum_i x_i \cdot P(X = x_i \cap Y = y_j)$$

$$= \sum_i \left[x_i \{\sum_j P(X = x_i \cap Y = y_j)\}\right]$$

$$= \sum x_i P(X = x_i) = E(X).$$

Theorem 6·14. *The variance of X can be regarded as consisting of two parts, the expectation of the conditional variance and the variance of the conditional expectation.* Symbolically,

$$V(X) = E[V(X \mid Y)] + V[E(X \mid Y)] \qquad \ldots(6\cdot51)$$

Proof. $E[V(X \mid Y)] + V[E(X \mid Y)]$

$$= E\Big[E(X^2 \mid Y) - \{E(X \mid Y)\}^2\Big] + E[\{E(X \mid Y)\}^2] - [E\{E(X \mid Y)\}]^2$$

$$= E[E(X^2 \mid Y)] - E[\{E(X \mid Y)\}^2] + E[\{E(X \mid Y)\}^2] - [E\{E(X \mid Y)\}]^2$$

$$= E[E(X^2 \mid Y)] - [E(X)]^2 \qquad (c.f.\ \text{Theorem } 6\cdot13)$$

$$= E[\sum_i x_i^2 P(X = x_i \mid Y = y_j)] - [E(X)]^2$$

$$= E\left[\sum_i x_i^2 \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}\right] - [E(X)]^2$$

$$= \sum_j \left\{\left[\sum_i x_i^2 \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}\right] P(Y = y_j)\right\} - [E(X)]^2$$

$$= \sum_i \left[x_i^2 \sum_j P(X = x_i \cap Y = y_j)\right] - [E(X)]^2$$

$$= \sum_i x_i^2 P(X = x_i) - [E(X)]^2$$

$$= E(X^2) - [E(X)]^2 = \text{Var}(X)$$

Hence the theorem.

Remarks The proofs of Theorems 6·13 and 6·14 for continous r.v.'s X and Y are left as an exercise to the reader.

Theorem 6·15. *Let A and B be two mutually exclusive events, then*

$$E(X \mid A \cup B) = \frac{P(A)\,E(X \mid A) + P(B)\,E(X \mid B)}{P(A \cup B)} \qquad \ldots(6\cdot52)$$

where by def.,

$$E(X \mid A) = \frac{1}{P(A)} \sum_{x_i \in A} x_i P(X = x_i)$$

Proof. $E(X \mid A \cup B) = \dfrac{1}{P(A \cup B)} \displaystyle\sum_{x_i \in A \cup B} x_i \; P(X = x_i)$

Since A and B are mutually exclusive events,

$$\sum_{x_i \in A \cup B} x_i P(X = x_i) = \sum_{x_i \in A} x_i P(X = x_i) + \sum_{x_i \in B} x_i P(X = x_i)$$

$$\therefore \quad E(X \mid A \cup B) = \frac{1}{P(A \cup B)} [P(A)\,E(X \mid A) + P(B)\,E(X \mid B)]$$

Cor. $\qquad E(X) = P(A)\,E(X \mid A) + P(\bar{A})\,E(X \mid \bar{A}) \qquad \ldots(6\cdot53)$

The corollary follows by putting $B = \bar{A}$ in the above Theorem.

Example 6·31. *Two ideal dice are thrown. Let X_1 be the score on the first die and X_2 the score on the second die. Let Y denote the maximum of X_1 and X_2, i.e., $Y = \max(X_1, X_2)$.*

(i) Write down the joint distribution of Y and X_1,

(ii) Find the mean and variance of Y and co-variance (Y, X_1).

Solution. Each of the random variables X_1 and X_2 can take six values 1, 2, 3, 4, 5, 6 each with probability 1/6, *i.e.*,

$$P(X_1 = i) = P(X_2 = i) = 1/6 \ ; \quad i = 1, 2, 3, 4, 5, 6 \qquad ...(i)$$

$$Y = \text{Max}(X_1, X_2).$$

Obviously

$$P(X_1 = i, Y = j) = 0\,, \text{ if } j < i = 1, 2, ..., 6$$

$$P(X_1 = i, Y = i) = P(X_1 = i, X_2 \le i) = \sum_{j=1}^{i} P(X_1 = i, X_2 = j)$$

$$= \sum_{j=1}^{i} P(X_1 = i)\, P(X_2 = j) \qquad (\because X_1, X_2 \text{ are independent.})$$

$$= \sum_{j=1}^{i} \left(\frac{1}{36}\right) = \frac{i}{36} \ ; \quad i = 1, 2, ..., 6.$$

$$P(X_1 = i, Y = j) = P(X_1 = i, X_2 = j) \ ; \quad j > i$$

$$= P(X_1 = i)\, P(X_2 = j) = \frac{1}{36} \ ; \quad j > i = 1, 2, ..., 6.$$

The joint probability table of X_1 and Y is given as follows:

X_1 \ Y	1	2	3	4	5	6	*Marginal Totals*
1	1/36	1/36	1/36	1/36	1/36	1/36	6/36
2	0	2/36	1/36	1/36	1/36	1/36	6/36
3	0	0	3/36	1/36	1/36	1/36	6/36
4	0	0	0	4/36	1/36	1/36	6/36
5	0	0	0	0	5/36	1/36	6/36
6	0	0	0	0	0	6/36	6/36
Marginal Totals	1/36	3/36	5/36	7/36	9/36	11/36	1

$$E(Y) = 1.\frac{1}{36} + 2.\frac{3}{36} + 3.\frac{5}{36} + 4.\frac{7}{36} + 5.\frac{9}{36} + 6.\frac{11}{36}$$

$$= \frac{1}{36}\Big[1 + 6 + 15 + 28 + 45 + 66\Big] = \frac{161}{36}$$

$$E(Y^2) = 1^2.\frac{1}{36} + 2^2.\frac{3}{36} + 3^2.\frac{5}{36} + 4^2.\frac{7}{36} + 5^2.\frac{9}{36} + 6^2.\frac{11}{36} = \frac{791}{36}$$

$$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{791}{36} - \left(\frac{161}{36}\right)^2 = \frac{2555}{1296}$$

$$E(X_1) = \frac{6}{36}\left[1+2+3+4+5+6\right] = \frac{126}{36} = \frac{21}{6}$$

$$E(X_1 Y) = 1.\frac{1}{36} + 2.\frac{1}{36} + 3.\frac{1}{36} + 4.\frac{1}{36} + 5.\frac{1}{36} + 6.\frac{1}{36}$$
$$+ 4.\frac{2}{36} + 6.\frac{1}{36} + 8.\frac{1}{36} + 10.\frac{1}{36} + 12.\frac{1}{36}$$
$$+ 9.\frac{3}{36} + 12.\frac{1}{36} + 15.\frac{1}{36} + 18.\frac{1}{36}$$
$$+ 16.\frac{4}{36} + 20.\frac{1}{36} + 24.\frac{1}{36}$$
$$+ 25.\frac{5}{36} + 30.\frac{1}{36} + 36.\frac{6}{36}$$
$$= \frac{1}{36}\left[21+44+72+108+155+216\right] = \frac{1}{36} \times 616$$

$$\text{Cov}(X_1, Y) = E(X_1 Y) - E(X_1)E(Y)$$
$$= \frac{616}{36} - \frac{21}{6} \cdot \frac{161}{36} = \frac{3696-3381}{216} = \frac{315}{216}.$$

Example 6·32. *Let X and Y be two random variables each taking three values −1, 0 and 1, and having the joint probability distribution :*

(i) Show that X and Y have different expectations.

Y \ X	−1	0	1	*Total*
−1	0	·1	·1	·2
0	·2	·2	·2	·6
1	0	·1	·1	·2
Total	2	·4	·4	1·0

(ii) Prove that X and Y are uncorrelated.

(iii) Find Var X and Var Y.

(iv) Given that Y = 0, what is the conditional probability distribution of X ?

(v) Find $V(Y \mid X = -1)$.

Solution. *(i)* $E(Y) = \Sigma p_i y_i = -1(\cdot 2) + 0(\cdot 6) + 1(\cdot 2) = 0$

$E(X) = \Sigma p_i x_i = -1(\cdot 2) + 0(\cdot 4) + 1(\cdot 4) = \cdot 2$

$\therefore \quad E(X) \neq E(Y)$

(ii) $E(XY) = \Sigma p_{ij}\, y_i x_j$

$$= (-1)(-1)(0) + 0(-1)(\cdot 1) + 1(-1)(\cdot 1)$$
$$+ 0(-1)(\cdot 2) + 0(0)(\cdot 2) + 0(1)(\cdot 2)$$
$$+ 1(-1)(0) + 1(0)(\cdot 1) + 1(1)(\cdot 1)$$
$$= -0\cdot 1 + 0\cdot 1 = 0$$

$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0$

$\Rightarrow$ X and Y are uncorrelated (*c.f.* § 10·5)

(iii) $E(Y^2) = (-1)^2(\cdot 2) + 0(\cdot 6) + 1^2(\cdot 2) = \cdot 4$

$\therefore \quad V(Y) = E(Y^2) - [E(Y)]^2 = \cdot 4$

$E(X^2) = (-1)^2(\cdot 2) + 0(\cdot 4) + 1^2(\cdot 4) = \cdot 2 + \cdot 4 = \cdot 6$

$V(X) = \cdot 6 - \cdot 04 = \cdot 56$

(iv) $$P(X = -1 \mid Y = 0) = \frac{P(X = -1 \cap Y = 0)}{P(Y = 0)} = \frac{\cdot 2}{\cdot 6} = \frac{1}{3}$$

$$P(X = 0 \mid Y = 0) = \frac{P(X = 0 \cap Y = 0)}{P(Y = 0)} = \frac{\cdot 2}{\cdot 6} = \frac{1}{3}$$

$$P(X = 1 \mid Y = 0) = \frac{P(X = 1 \cap Y = 0)}{P(Y = 0)} = \frac{\cdot 2}{\cdot 6} = \frac{1}{3}$$

(v) $$V(Y \mid X = -1) = E(Y \mid X = -1)^2 - \{E(Y \mid X = -1)\}^2$$

$$E(Y \mid X = -1) = \sum_y y\, P(Y = y \mid X = -1) = (-1)0 + 0(\cdot 2) + 1(0) = 0$$

$$E(Y \mid X = -1)^2 = \sum_y y^2 P(Y = y \mid X = -1) = 1(0) + 0(\cdot 2) + 1(0) = 0$$

$\therefore \quad V(Y \mid X = -1) = 0.$

Example 6·33. *Two tetrahedra with sides numbered 1 to 4 are tossed. Let X denote the number on the downturned face of the first tetrahedron and Y denote the larger of the downturned numbers. Investigate the following :*

(a) Joint density function of X, Y and marginals f_X and f_Y,

(b) $P\{X \le 2, Y \le 3\}$, *(c) $\rho(X, Y)$,* *(d) $E(Y \mid X = 2)$,*

(e) Construct joint density different from that in part (a) but possessing same marginals f_X and f_Y. **[Delhi Univ. B.A. (Stat. Hons.), Spl. Course, 1985]**

Hint. The sample space is $S = \{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$ and each of the 16 sample points (outcomes) has probability $p = 1/16$ of occurrence.

Let X : Number on the first dice and Y : Larger of the numbers on the two dice. Then the above 16 sample points, in that order, give the following distribution of X and Y.

Sample Point	:	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(2, 1)	(2, 2)	(2, 3)	(2, 4)
X	:	1	1	1	1	2	2	2	2
Y	:	1	2	3	4	2	2	3	4
Sample Point	:	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(4, 1)	(4, 2)	(4, 3)	(4, 4)
X	:	3	3	3	3	4	4	4	4
Y	:	3	3	3	4	4	4	4	4

Since each sample point has probability $p = 1/16$, the joint density functions of X and Y and the marginal densities f_X and f_Y are given on page 6·61.

Here $p = 1/16$.

(b) $P(X \le 2, Y \le 3) = p + p + 2p + p + p = 6p = 3/8.$

(c) $\text{Var}(X) = EX^2 - [E(X)]^2 = \frac{15}{2} - \frac{25}{4} = \frac{5}{4}$ (Try it)

(a)	x = 1	2	3	4	Total (f_y)
y = 1	p	0	0	0	p
2	p	$2p$	0	0	$3p$
3	p	p	$3p$	0	$5p$
4	p	p	p	$4p$	$7p$
Total (f_x)	$4p$	$4p$	$4p$	$4p$	$16p=1$

(e)	x = 1	2	3	4	Total (f_y)
y = 1	p	0	0	0	p
2	p	$2p$	0	0	$3p$
3	p	$p+\epsilon$	$3p-\epsilon$	0	$5p$
4	p	$p-\epsilon$	$p+\epsilon$	$4p$	$7p$
Total (f_x)	$4p$	$4p$	$4p$	$4p$	1

$$\text{Var}(Y) = EY^2 - [E(Y)]^2 = \frac{85}{8} - \left(\frac{25}{8}\right)^2 = \frac{55}{64} \qquad \text{(Try it)}$$

$$\text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = \frac{135}{16} - \frac{5}{2} \times \frac{25}{8} = \frac{5}{8} \qquad \text{(Try it)}$$

$$\therefore \quad \rho(X, Y) = \frac{5/8}{\sqrt{5/4 \times 55/64}} = \frac{2}{\sqrt{11}}$$

$$(d) \quad E(Y \mid X = 2) = \Sigma y \,.\, f(y \mid x = 2) = \Sigma y \,.\, \frac{f(x = 2 \cap y)}{f(x = 2)}$$

$$= 4 \,.\, \Sigma y\, f(2, y) = 4\,[0 + 4p + 3p + 4p] = 44p = \frac{11}{4}$$

(e) Let $0 < \epsilon < p$. The joint density of X *and* Y given in (*e*) above is different from that in (*a*) but has the same marginals as in (*a*).

Example 6·34. *(a) Given two variates X_1 and X_2 with joint density function $f(x_1, x_2)$, prove that conditional mean of X_2 (given X_1) coincides with (unconditional) mean only if X_1 and X_2 are independent (stochastically).*

(b) Let $f(x_1, x_2) = 21x_1^2 x_2^3,\ 0 < x_1 < x_2 < 1,$ and zero elsewhere be the joint p.d.f. of X_1 and X_2. Find the conditional mean and variance of X_1 given $X_2 = x_2$, $0 < x_2 < 1$. **[Delhi Univ. M.A. (Eco.), 1986]**

Solution. *(a)* Conditional mean of X_2 given X_1 is given by :

$$E(X_2 \mid X_1 = x_1) = \int_{x_2} x_2 f(x_2 \mid x_1)\, dx_2 \qquad \ldots(*)$$

where $f(x_2 \mid x_1)$ is conditional p.d.f. of X_2 given $X_1 = x_1$.

But the joint p.d.f. of X_1 and X_2 is given by

$$f(x_1, x_2) = f_1(x_1) \,.\, f(x_2 \mid x_1)$$

$$\Rightarrow \qquad f(x_2 \mid x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}$$

where $f_1(\cdot)$ is marginal p.d.f. of X_1.

Substituting in (*), we get

$$E(X_2 \mid X_1 = x_1) = \int_{x_2} \left[\frac{x_2 f(x_1, x_2)}{f_1(x_1)} \right] dx_2, \qquad \text{...(**)}$$

Unconditional mean of X_2 is given by

$$E(X_2) = \int_{x_2} x_2 \, f_2(x_2) \, dx_2, \qquad \text{...(***)}$$

where $f_2(\cdot)$ is marginal p.d.f. of X_2.

From (**) and (***), we conclude that the conditional mean of X_2 (given X_1) will coincide with unconditional mean of X_2 only if

$$\frac{f(x_1, x_2)}{f_1(x_1)} = f_2(x_2)$$

$$\Rightarrow \quad f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$$

i.e., if X_1 and X_2 are (stochastically) independent.

(b) $\quad f(x_1, x_2) = 21\, x_1^2 x_2^3 \,;\ 0 < x_1 < x_2 < 1$
$\quad = 0, \quad$ otherwise

Marginal p.d.f. of X_2 is given by

$$f_2(x_2) = \int_0^{x_2} f(x_1, x_2)\, dx_1 = 21\, x_2^3 \int_0^{x_2} x_1^2 \, dx_1$$

$$= 21\, x_2^3 \left| \frac{x_1^3}{3} \right|_0^{x_2} = 7\, x_2^6 \,;\quad 0 < x_2 < 1$$

$\therefore$ Conditional p.d.f. of X_1 (given X_2) is given by

$$f_1(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} = 3\, \frac{x_1^2}{x_2^3} \,;\ 0 < x_1 < x_2 \,;\, 0 < x_2 < 1$$

Conditional mean of X_1 is

$$E(X_1 \mid X_2 = x_2) = \int_0^{x_2} x_1 f_1(x_1 \mid x_2)\, dx_1 = \frac{3}{x_2^3} \int_0^{x_2} x_1^3 \, dx_1$$

$$= \frac{3}{x_2^3} \cdot \left| \frac{x_1^4}{4} \right|_0^{x_2} = \frac{3x_2}{4} \,;\ 0 < x_2 < 1$$

Now

$$E(X_1^2 \mid X_2 = x_2) = \int_0^{x_2} x_1^2 f_1(x_1 \mid x_2)\, dx_1 = \frac{3}{x_2^3} \int_0^{x_2} x_1^4 \, dx_1$$

$$= \frac{3}{x_2^3} \cdot \frac{x_2^5}{5} = \frac{3}{5} x_2^2$$

$$\therefore \quad \text{Var}(X_1 \mid X_2 = x_2) = E(X_1^2 \mid X_2 = x_2) - [E(X_1 \mid X_2 = x_2)]^2$$

$$= \frac{3}{5} x_2^2 - \frac{9}{16} x_2^2 = \frac{3}{80} x_2^2 \,;\ 0 < x_2 < 1.$$

Example 6·35. *Two random variables X and Y have the followinig joint probability density function :*

$$f(x, y) = 2 - x - y\ ;\ 0 \le x \le 1,\ 0 \le y \le 1$$
$$= 0,\ otherwise$$

Find (i) *Marginal probability density functions of X and Y.*
(ii) *Conditional density functions.*
(iii) *Var (X) and Var (Y).*
(iv) *Co-variance between X and Y.*

[Dibrugarh Univ. B.Sc. (Hons.), 1991]

Solution. (i) $f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy$

$$= \int_0^1 (2 - x - y)\, dy = \frac{3}{2} - x$$

$\therefore$ $$f_X(x) = \frac{3}{2} - x,\ 0 < x < 1$$
$$= 0,\ \text{otherwise}$$

Similarly $$f_Y(y) = \frac{3}{2} - y,\ 0 < y < 1$$
$$= 0,\ \text{otherwise}$$

(ii) $$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{(2 - x - y)}{(3/2 - y)},\ 0 < (x, y) < 1$$

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{(2 - x - y)}{(3/2 - x)},\ 0 < (x, y) < 1$$

$$E(X) = \int_0^1 x f_X(x)\, dx = \int_0^1 x\left(\frac{3}{2} - x\right) dx = \frac{5}{12}$$

$$E(Y) = \int_0^1 y f_Y(y)\, dy = \int_0^1 y\left(\frac{3}{2} - y\right) dy = \frac{5}{12}$$

(iii) $$E(X^2) = \int_0^1 x^2\left(\frac{3}{2} - x\right) dx = \left[\frac{3}{6}x^3 - \frac{x^4}{4}\right]_0^1 = \frac{1}{4}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{1}{4} - \frac{25}{144} = \frac{11}{144}$$

Similarly $$V(Y) = \frac{11}{144}$$

(iv) $$E(XY) = \int_0^1\int_0^1 xy(2 - x - y)\, dx\, dy$$

$$= \int_0^1 \left| 2\frac{x^2y}{2} - \frac{x^3y}{3} - \frac{x^2y^2}{2} \right|_0^1 dy$$

$$= \int_0^1 \left(\frac{2}{3}y - \frac{1}{2}y^2\right) dy$$

$$= \left| \frac{y^2}{3} - \frac{y^3}{6} \right|_0^1 = \frac{1}{6}$$

$$= \left| \frac{y^2}{3} - \frac{y^3}{6} \right|_0^1 = \frac{1}{6}$$

$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{1}{6} - \frac{5}{12} \cdot \frac{5}{12} = -\frac{1}{144}.$

Example 6·36. *Let* $f(x, y) = 8xy, 0 < x < y < 1$; $f(x, y) = 0$, *elsewhere. Find* (a) $E(Y \mid X = x)$, (b) $E(XY \mid X = x)$, (c) $Var(Y \mid X = x)$. **[Calcutta Univ. B.Sc. (Maths Hons.), 1988; Delhi Univ. B.Sc. (Maths Hons.), 1990]**

Solution. $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$

$$= 8x \int_x^1 y\, dy$$

$$= 4x(1 - x^2), 0 < x < 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

$$= 8y \int_0^y x\, dx$$

$$= 4y^3, 0 < y < 1$$

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)} = \frac{2x}{y^2}, \; f_{Y|X}(y \mid x) = \frac{2y}{1 - x^2}, 0 < x < y < 1.$$

(a) $$E(Y \mid X = x) = \int_x^1 y \left(\frac{2y}{1 - x^2} \right) dy = \frac{2}{3} \left(\frac{1 - x^3}{1 - x^2} \right) = \frac{2}{3} \left(\frac{1 + x + x^2}{1 + x} \right)$$

(b) $$E(XY \mid X = x) = x\, E(Y \mid X = x) = \frac{2}{3} \cdot \frac{x(1 + x + x^2)}{(1 + x)}$$

(c) $$E(Y^2 \mid X = x) = \int_x^1 y^2 \left(\frac{2y}{1 - x^2} \right) dy = \frac{1}{2} \left(\frac{1 - x^4}{1 - x^2} \right) = \frac{1 + x^2}{2}$$

$$\text{Var}(Y \mid X = x) = E(Y^2 \mid X = x) - [E(Y \mid X = x)]^2$$

$$= \frac{1 + x^2}{2} - \frac{4}{9} \cdot \frac{(1 + x + x^2)^2}{(1 + x)^2}$$

EXERCISE 6(b)

1. The joint probability distribution of X and Y is given by the following table:

X \ Y	1	3	9
2	1/8	1/24	1/12
4	1/4	1/4	0
6	1/8	1/24	1/12

(i) Find the marginal probability distribution of Y.

(ii) Find the conditional distribution of Y given that $X = 2$,

(iii) Find the covariance of X and Y, and

(iv) Are X and Y independent ?

2. A fair coin is tossed four times. Let X denote the number of heads occurring and let Y denote the longest string of heads occurring.

(i) Determine the joint distribution of X and Y, *and* *(ii)* Find Cov (X, Y).

Hint.

X \ Y	0	1	2	3	4	*Total*
0	1/16	0	0	0	0	1/16
1	0	4/16	0	0	0	4/16
2	0	3/16	3/16	0	0	6/16
3	0	0	2/16	2/16	0	4/16
4	0	0	0	0	1/16	1/16
Total	1/16	7/16	5/16	2/16	1/16	1

(ii) Cov $(X, Y) = 0·85$.

3. X and Y are jointly discrete random variables with probability function

$$p(x, y) = 1/4 \text{ at } (x, y) = (-3, -5), (-1, -1), (1, 1), (3, 5)$$
$$= 0, \text{ otherwise}$$

Compute $E(X)$, $E(Y)$, $E(XY)$ and $E(X \mid Y)$. Are X and Y independent ?

4. X_1 and X_2 have a bivariate distribution given by

$$P(X_1 = x_1 \cap X_2 = x_2) = \frac{x_1 + 3x_2}{24}, \text{ where } (x_1, x_2) = (1, 1), (1, 2), (2, 1), (2, 2)$$

Find the conditional mean and variance of X_1, given $X_2 = 2$.

5. Two random variables X and Y have the following joint probability density function :

$$f(x, y) = k(4 - x - y) \ ; \ 0 \le x \le 2 \ ; \ 0 \le y \le 2$$
$$= 0, \text{ otherwise}$$

Find *(i)* the constant k,

(ii) marginal density functions of X and Y,

(iii) conditional density functions, and

(iv) Var (X), Var (Y) and Cov (X, Y). (Poona Univ. B.Sc., Oct. 1991)

6. Let the joint probability density function of the random variables X and Y be

$$f(x, y) = 2(x + y - 3xy^2) \ ; \ 0 < x < 1, 0 < y < 1$$
$$= 0, \text{ otherwise}$$

(i) Find the marginal distributions of X and Y.

(ii) Is $E(XY) = E(X)\,E(Y)$?

(*iii*) Find $E(X+Y)$ and $E(X-Y)$. **[Calicut Univ. B.Sc., Oct. 1990]**

7. (*a*) Let X and Y have the joint probability density function

$$f(x, y) = 2, 0 < x < y < 1$$
$$= 0, \text{ otherwise}$$

Show that the conditional mean and variance of X given $Y = y$ are $y/2$ and $y^2/12$ respectively.

(*b*) If $f(x, y) = 2$; $0 < x < y, 0 < y < 1$

Find (*i*) $E(Y \mid X)$, (*ii*) $E(X \mid Y)$.

8. Give an example to show that $E(Y)$ may not exist though $E(XY)$ and $E(Y \mid X)$ may both exist? **[Delhi Univ. B.A. (Stat. Hons.) Spl. Course, 1985]**

Hint. Consider the joint p.d.f. :

$$f(x, y) = x \cdot e^{-x(1+y)}; x \geq 0, y \geq 0.$$
$$= 0, \text{ otherwise.}$$

Then we shall get :

$$f_X(x) = \int_0^\infty f(x, y)\, dy = e^{-x}; x \geq 0$$

$$f_Y(y) = \int_0^\infty f(x, y)\, dx = \frac{1}{(1+y)^2}; \; y \geq 0$$

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} = x\, e^{-xy}; \; y \geq 0$$

$$\therefore \quad E(Y) = \int_0^\infty y f(y) = \infty \quad \Rightarrow \quad E(Y) \text{ does not exist.}$$

$$E(XY) = \int_0^\infty \int_0^\infty xy \cdot f(x\, y)\, dx\, dy$$

$$E(Y \mid X = x) = \int_0^\infty y \cdot f(y \mid x)\, dy = \frac{1}{x}$$

$\Rightarrow$ Both $E(XY)$ and $(E(Y \mid X = x)$ exist, though $E(Y)$ does not exist.

9. Three coins are tossed. Let X denote the number of heads on the first two coins, Y denote the number of tails on the last two and Z denote the number of heads on the last two.

(*a*) Find the joint distribution of (*i*) X and Y, (*ii*) X and Z.

(*b*) Find the conditional distribution of Y given $X = 1$.

(*c*) Find $E(Z \mid X = 1)$.

(*d*) Find $\rho_{X, Y}$ and $\rho_{X, Z}$.

(*e*) Give a joint distribution that is not the joint distribution of X and Z in (*a*) and yet has the same marginals as $f(x, z)$ has in part (*a*).

[Delhi Univ. B.Sc. (Maths Hons.), 1989]

Hint. The sample space is $S' = \{H, T\} \times \{H, T\} \times \{H, T\}$

$= \{H, T\} \times \{HH, HT, TH, TT\}$

and each of the 8 sample points (outcomes) has the probability $p = 1/8$ of occurrence.

X : Number of heads on the 1st two coins.

Y : Number of tails on the last two coins.

Z : Number of heads on the last two coins.

Then the distribution of X, Y and Z is given below :

Sample Point :	*HHH*	*HHT*	*HTH*	*HTT*	*THH*	*THT*	*TTH*	*TTT*
Probability	p	p	p	p	p	p	p	p
X	2	2	1	1	1	1	0	0
Y	0	1	1	2	0	1	1	2
Z	2	1	1	0	2	1	1	0

Joint Distribution of X and Y

		y			*Total*
		0	1	2	(f_x)
x	0	0	1/8	1/8	1/4
	1	1/8	2/8	1/8	1/2
	2	1/8	1/8	0	1/4
Total (f_y)		1/4	1/2	1/4	1

Joint Distribution of X and Z

		z			*Total*
		0	1	2	(f_x)
x	0	1/8	1/8	0	1/4
	1	1/8	2/8	1/8	1/2
	2	0	1/8	1/8	1/4
Total (f_z)		1/4	1/2	1/4	1

(*b*) $P(Y=0 \mid X=1) = \dfrac{P(Y=0,\ X=1)}{P(X=1)} = \dfrac{1/8}{1/2} = \dfrac{1}{4}$

Similarly, $P(Y=1 \mid X=1) = \dfrac{2/8}{1/2} = \dfrac{1}{2}$; $P(Y=2 \mid X=1) = \dfrac{1/8}{1/2} = \dfrac{1}{4}$

(*c*) $E(Z \mid X=1) = \Sigma z \, . \, P(Z \mid X=1) = 0 \times \dfrac{1/8}{1/2} + 1 \times \dfrac{2/8}{1/2} + 2 \times \dfrac{1/8}{1/2} = 1$

(*d*) $\rho_{XY} = \dfrac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \dfrac{-1/4}{\sqrt{1/2 \times 1/2}} = -\dfrac{1}{2}$.

$\rho_{XZ} = \dfrac{\text{Cov}(X, Z)}{\sigma_X \sigma_Z} = \dfrac{-1/4}{\sqrt{1/2 \times 1/2}} = -\dfrac{1}{2}$

(*e*) Let $0 \le \varepsilon \le 1/8$. The joint probability distribution of (X, Z) given belc has the same marginals as in part (*a*).

		0	Z 1	2	Total (f_x)
X	0	1/8	1/8	0	1/4
	1	1/8	$2/8 + \varepsilon$	$1/8 - \varepsilon$	1/2
	2	0	$1/8 - \varepsilon$	$1/8 + \varepsilon$	1/4
Total (f_z)		1/4	1/2	1/4	1

10. Let $f_{XY}(x, y) = e^{-(x+y)}$; $0 < x < \infty, 0 < y < \infty$

Find :

(a) $P(X > 1)$ (d) m so that $P(X + Y < m) = 1/2$

(b) $P(1 < X + Y < 2)$ (e) $P(0 < X < 1 \mid Y = 2)$

(c) $P(X < Y \mid X < 2Y)$ (f) ρ_{XY}

Ans. $f_X(x) = e^{-x}$; $x \geq 0$; $f_Y(y) = e^{-y}$; $y \geq 0$

(a) $1/e$ (b) *Hint.* $X + Y$ is a Gamma variabte with parameter $n = 2$. [See Chapter 8] $(2/e - 3/e^2)$.

$$(c)\ P(X < Y \mid X < 2Y) = \frac{P(X < Y \cap X < 2Y)}{P(X < 2Y)} = \frac{P(X < Y)}{P(X < 2Y)} = \frac{1/2}{2/3} = \frac{3}{4}$$

(d) Use hint in (b). $e^{-m}(1 + m) = 1/2$; (e) $(e - 1)/e$

(f) $f_{XY}(x, y) = f_X(x) f_Y(y) \Rightarrow X$ and Y are independent $\Rightarrow \rho_{XY} = 0$.

11. The joint p.d.f. of X and Y is given by :

$$f(x, y) = 3(x + y) \ ; \ 0 \leq x \leq 1, \ 0 \leq y \leq 1 ; \ 0 \leq x + y \leq 1$$

Find : (a) Marginal density of X. (b) $P(X + Y < 1/2)$

(c) $E(Y \mid X = x)$ (d) Cov (X, Y)

Ans. (a) $f_X(x) = \frac{3}{2}(1 - x^2)$; $0 \leq x \leq 1$.

$$(b)\ P(X + Y < 1/2) = \int_0^{0.5}\left[\int_0^{0.5-x} 3(x + y)\, dy\right] dx = \frac{1}{8}$$

$$(c)\ \frac{(1 - x)(x + 2)}{3(1 + x)} \qquad (d)\ E(XY) = \int_0^1\left[\int_0^{1-x} xy\ f(x, y) dy\ dx = \frac{1}{10}\right]$$

$$\text{Cov}(x, y) = E(XY) - E(X)E(Y) = \frac{1}{10} - \frac{3}{8} \times \frac{3}{8} = -\frac{13}{320}.$$

6·10. Moment Generating Function. The moment generating function (m.g.f.) of a random variable X (about origin) having the probability function $f(x)$ is given by

$$(6 \cdot 54) \ldots \quad \begin{cases} M_X(t) = E(e^{tX}) = \int e^{tx} f(x)\, dx, & \text{(for } continuous\ probability\ distribution\text{)} \\ = \sum_x e^{tx} f(x), & \text{(for } discrete\ probability\ distribution\text{)} \end{cases}$$

the integration or summation being extended to the entire range of x, t being the real parameter and it is being assumed that the right-hand side of (6·54) is absolutely convergent for some positive number h such that $-h < t < h$. Thus

$$M_X(t) = E(e^{tX}) = E\left[1 + tX + \frac{t^2X^2}{2!} + \ldots + \frac{t^rX^r}{r!} + \ldots\right]$$

$$= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \ldots + \frac{t^r}{r!}E(X^r) + \ldots$$

$$= 1 + t\,\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots + \frac{t^r}{r!}\mu_r' + \ldots \qquad \ldots(6{\cdot}55)$$

where $\mu_r' = E(X^r) = \int x^r f(x)\,dx$, *for continuous distribution*

$= \sum_x x^r p(x)$, *for discrete distribution*,

is the rth moment of X about origin. Thus we see that the coefficient of $\frac{t^r}{r!}$ in $M_X(t)$ gives μ_r' (above origin). Since $M_X(t)$ generates moments, it is known as moment generating function.

Differentiating (6·55) w.r.t. t and then putting $t = 0$, we get

$$\left[\frac{d^r}{dt^r}\{M_X(t)\}\right]_{t=0} = \left[\frac{\mu_r'}{r!}\cdot r! + \mu'_{r+1}\,t + \mu'_{r+2}\cdot\frac{t^2}{2!} + \ldots\right]_{t=0}$$

$$\Rightarrow \qquad \mu_r' = \left[\frac{d^r}{dt^r}\{M_X(t)\}\right]_{t=0} \qquad \ldots(6{\cdot}56)$$

In general, the moment generating function of X about the point $X = a$ is defined as

$$M_X(t)\ (\text{about } X = a) = E[e^{t(X-a)}]$$

$$= E\left[1 + t(X-a) + \frac{t^2}{2!}(X-a)^2 + \ldots + \frac{t^r}{r!}(X-a)^r + \ldots\right]$$

$$= 1 + t\,\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots + \frac{t^r}{r!}\mu_r' + \ldots \qquad \ldots(6{\cdot}57)$$

where $\mu_r' = E\{(X-a)^r\}$, is the rth moment about the point $X = a$.

6·10·1. Some Limitations of Moment Generating Functions. Moment generating function suffers from some drawbacks which have restricted its use in Statistics. We give below some of the deficiencies of m.g.f.'s with illustrative examples.

1. *A random variable X may have no moments although its m.g.f. exists.* For example, let us consider a discrete r.v. with probability function

$$f(x) = \begin{cases} \dfrac{1}{x(x+1)}; & x = 1, 2, \ldots \\ = 0, & \text{otherwise} \end{cases}$$

Here $$E(X) = \sum_{x=1}^{\infty} x f(x) = \sum_{x=1}^{\infty} \frac{1}{(x+1)}$$

$$= \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$$

$$= \left[\sum_{x=1}^{\infty} \frac{1}{x}\right] - 1$$

Since $\sum_{x=1}^{\infty} \frac{1}{x}$ is a divergent series, $E(X)$ does not exist and consequently no moment of X exists. However, the m.g.f. of X is given by

$$
\begin{aligned}
M_X(t) &= \sum_{x=1}^{\infty} e^{tx} . f(x) = \sum_{x=1}^{\infty} \frac{e^{tx}}{x(x+1)} \\
&= \sum_{x=1}^{\infty} \frac{z^x}{x(x+1)}, \quad (z = e^t) \qquad \ldots(*) \\
&= \frac{z}{1.2} + \frac{z^2}{2.3} + \frac{z^3}{3.4} + \frac{z^4}{4.5} + \ldots \\
&= z\left[1 - \frac{1}{2}\right] + z^2\left[\frac{1}{2} - \frac{1}{3}\right] + z^3\left[\frac{1}{3} - \frac{1}{4}\right] + z^4\left[\frac{1}{4} - \frac{1}{5}\right] + \ldots \\
&= \left[z + \frac{z^2}{2} + \frac{z^3}{3} + \frac{z^4}{4} + \ldots\right] - \frac{z}{2} - \frac{z^2}{3} - \frac{z^3}{4} - \frac{z^4}{5} - \ldots \\
&= -\log(1-z) - \frac{1}{z}\left[\frac{z^2}{2} + \frac{z^3}{3} + \frac{z^4}{4} + \ldots\right] \\
&= -\log(1-z) + 1 + \frac{1}{z}\log(1-z), \ |z| < 1 \\
&= 1 + \left[\frac{1}{z} - 1\right]\log(1-z), \ |z| < 1 \\
&= 1 + (e^{-t} - 1)\log(1 - e^t), \ t < 0
\end{aligned}
$$

$$[\because |z| < 1 \Rightarrow |e^t| < 1 \Rightarrow t < 0]$$

And $\quad M_X(t) = 1,$ for $t = 0,$ [From (*)]

while for $t > 0$, $M_X(t)$ does not exist.

2. *A random variable X can have m.g.f. and some (or all) moments, yet the m.g.f. does not generate the moments.* For example, consider a discrete r.v. with probability function

$$P(X = 2^x) = \frac{e^{-1}}{x!}\ ; \ x = 0, 1, 2, \ldots$$

Here
$$E(X^r) = \sum_{\xi=0}^{\infty} (2^x)^r P(X = 2^x) = e^{-1} \sum_{\xi=0}^{\infty} \frac{(2^r)^x}{x!}$$
$$= e^{-1} . \exp(2^r) = \exp(2^r - 1)$$

Hence all the moments of X exist.

The m.g.f. of X, if it exists is given by

$$M_X(t) = \sum_{x=0}^{\infty} \exp(t . 2^x)\left(\frac{e^{-1}}{x!}\right) = e^{-1} \sum_{x=0}^{\infty} \exp(t . 2^x)\frac{1}{x!}$$

By D'Alembert's ratio test, the series on the R.H.S. converges for $t \leq 0$ and diverges for $t > 0$. Hence $M_X(t)$ cannot be differentiated at $t = 0$ and has no Maclaurin's expansion and consequently it does not generate moments.

3. *A r.v. X can have all or some moments; but m.g.f. does not exist except perhaps at one point.*

For example, let X be a r.v. with probability function

$$P(X = \pm 2^x) = \frac{e^{-1}}{2x!} \; ; \; x = 0, 1, 2, \ldots$$
$$= 0, \text{ otherwise.}$$

Since the distribution is symmetric about the line $X = 0$, all moments of odd order about origin vanish, *i.e.*,

$$E(X^{2r+1}) = 0 \Rightarrow \mu_{2r+1} = 0$$

$$E(X^{2r}) = \sum_{x=0}^{\infty} (\pm 2^x)^{2r} \left(\frac{1}{2e x!}\right) = \frac{1}{e} \sum_{x=0}^{\infty} \frac{(2^{2r})^x}{x!}$$

$$= \frac{1}{e} \cdot \exp(2^{2r}) = \exp(2^{2r} - 1)$$

Thus all the moments of X exist. The m.g.f. of X, if it exists, is given by

$$M_X(t) = \sum_{x=0}^{\infty} \left[\left(e^{t.2^x} + e^{-t.2^x}\right) \frac{1}{2e\, x!} \right]$$

$$= e^{-1} \sum_{x=0}^{\infty} \left[\frac{\cos(t.2^x)}{x!} \right]$$

which converges only for $t = 0$.

As an illustration of a continuous probability distribution, consider Pareto distribution with p.d.f.

$$p(x) = \frac{\theta \,.\, a^{\theta}}{x^{\theta+1}} \; ; \quad x \geq a \; ; \; \theta > 1$$

$$E(X^r) = \theta \,.\, a^{\theta} \int_a^{\infty} x^{r-\theta-1}\, dx = \theta \,.\, a^{\theta} \,.\, \left| \frac{x^{r-\theta}}{r-\theta} \right|_a^{\infty},$$

which is finite iff $r - \theta < 0 \Rightarrow \theta > r$ and then

$$E(X^r) = \theta\, a^{\theta} \left[0 - \frac{a^{r-\theta}}{r-\theta} \right] = \frac{\theta \,.\, a^r}{\theta - r} \; ; \quad \theta > r$$

However, the m.g.f. is given by :

$$M_X(t) = \theta \,.\, a^{\theta} \int_a^{\infty} \frac{e^{tx}}{x^{\theta+1}}\, dx,$$

which does not exist, since e^{tx} dominates $x^{\theta+1}$ and $(e^{tx}/x^{\theta+1}) \to \infty$ as $x \to \infty$ and hence the integral is not convergent.

For more illustrations see Student's t-distribution and Snedecor's F-distributions, for which m.g.f.'s do not exist, though the moments of all orders exist. [c.f. Chapter 14, § 14·2·4 and 14·5·2.].

Remark. The reason that m.g.f. is a poor tool in comparison with characteristic function (c.f. § 6·12) is that the domain of the dummy parameter 't' of the m.g.f. depends on the distribution of the r.v. under consideration, while characteristic function exists for all real t, $(-\infty < t < \infty)$. If m.g.f. is valid for t lying in an interval containing zero, then m.g.f. can be expanded with perhaps some additional restrictions.

6·10·2. Theorems on Moment Generating Functions.

Theorem 6·17. $M_{cX}(t) = M_X(ct)$, c being a constant. ...(6·58)

Proof. By def.,

$$\text{L.H.S.} = M_{cX}(t) = E(e^{t.cX})$$

$$\text{R.H.S.} = M_X(ct) = E(e^{ctX}) = \text{L.H.S.}$$

Theorem 6·18. *The moment generating function of the sum of a number of independent random variables is equal to the product of their respective moment generating functions.*

Symbolically, if $X_1, X_2, ..., X_n$ are independent random variables, then the moment generating function of their sum $X_1 + X_2 + ... + X_n$ is given by

$$M_{X_1+X_2+...+X_n}(t) = M_{X_1}(t)\, M_{X_2}(t) \dots M_{X_n}(t) \qquad ...(6·59)$$

Proof. By definition,

$$M_{X_1+X_2+...+X_n}(t) = E\left[e^{t(X_1+X_2+...+X_n)}\right]$$

$$= E\left[e^{tX_1} . e^{tX_2} \dots e^{tX_n}\right]$$

$$= E(e^{tX_1})\, E(e^{tX_2}) \dots E(e^{tX_n})$$

$$(\because X_1, X_2, ..., X_n \text{ are independent})$$

$$= M_{X_1}(t) . M_{X_2}(t) \dots M_{X_n}(t)$$

Hence the theorem.

Theorem 6·19. *Effect of change of origin and scale on M.G.F.* Let us transform X to the new variable U by changing both the origin and scale in X as follows :

$$U = \frac{X-a}{h}, \text{ where } a \text{ and } h \text{ are constants}$$

M.G.F. of U (about origin) is given by

$$M_U(t) = E(e^{tU}) = E\left[\exp\{t(x-a)/h\}\right]$$

$$= E\left[e^{tX/h} . e^{-at/h}\right] = e^{-at/h} E(e^{tX/h})$$

$$= e^{-at/h} E(e^{Xt/h}) = e^{-at/h} M_X(t/h) \qquad ...(6·60)$$

where $M_X(t)$ is the m.g.f. of X about origin.

In particular, if we take $a = E(X) = \mu$ (say) and $h = \sigma_X = \sigma$ (say), then

$$U = \frac{X - E(X)}{\sigma_X} = \frac{X-\mu}{\sigma} = Z \text{ (say)},$$

is known as a *standard variate.* Thus the m.g.f. of a standard variate Z is given by

$$M_Z(t) = e^{-\mu t/\sigma} M_X(t/\sigma) \qquad ...(6·61)$$

Remark. $E(Z) = E\left(\dfrac{X-\mu}{\sigma}\right) = \dfrac{1}{\sigma} E(X-\mu)$

$$= \frac{1}{\sigma}\{E(X) - \mu\} = \frac{1}{\sigma}(\mu - \mu) = 0$$

and $\quad V(Z) = V\left(\dfrac{X-\mu}{\sigma}\right) = \dfrac{1}{\sigma^2} V(X-\mu)$ [c.f. Cor. (i) Theorem 6·8]

$= \dfrac{1}{\sigma^2} V(X) = \dfrac{1}{\sigma}\sigma^2 = 1$ [c.f. Cor. (iii) Theorem 6·8]

$\therefore \quad E(Z) = 0$ and $V(Z) = 1$, *i.e.*, the mean and variance of a standard variate are 0 and 1 respectively.

6·10·3. Uniqueness Theorem of Moment Generating Function. *The moment generating function of a distribution, if it exists, uniquelly determines the distribution.* This implies that corresponding to a given probability distribution, there is only one m.g.f. (provided it exists) and corresponding to a given m.g.f., there is only one probability distribution. Hence $M_X(t) = M_Y(t) \Rightarrow X$ and Y are identically distributed. [For detailed discussion, see Uniqueness Theorem of Characteristic Functions – Theorem 6·27, page 6·90]

6·11. Cumulants. Cumulants generating function $K(t)$ is defined as

$$K_X(t) = \log_e M_X(t), \qquad \ldots(6\text{·}62)$$

provided the right-hand side can be expanded as a convergent series in powers of t. Thus

$$K_X(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \ldots + \kappa_r \frac{t^r}{r!} + \ldots = \log M_X(t)$$

$$= \log\left[1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \ldots + \mu_r' \frac{t^r}{r!} + \ldots\right] \qquad \ldots(6\text{·}62a)$$

where κ_r = coefficient of $\dfrac{t^r}{r!}$ in $K_X(t)$ is called the rth cumulant. Hence

$$\kappa_1 t + \kappa_2 \frac{t^2}{2!} + \kappa_3 \frac{t^3}{3!} + \kappa_4 \frac{t^4}{4!} + \ldots$$

$$= \left[\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \mu_4' \frac{t^4}{4!} + \ldots\right)\right.$$

$$- \frac{1}{2}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \ldots\right)^2$$

$$\left. + \frac{1}{3}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \ldots\right)^3 - \frac{1}{4}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \ldots\right)^4 + \ldots\right]$$

Comparing the coefficients of like powers of 't' on both sides, we get the relationship between the moments and cumulants. Hence, we have

$$\kappa_1 = \mu_1' = \text{Mean}, \quad \frac{\kappa_2}{2!} = \frac{\mu_2'}{2!} - \frac{\mu_1'^2}{2!} \quad \Rightarrow \quad \kappa_2 = \mu_2' - \mu_1'^2 = \mu_2$$

$$\frac{\kappa_3}{3!} = \frac{\mu_3'}{3!} - \frac{1}{2}\frac{2\mu_1'\mu_2'}{2!} + \frac{\mu_1'^3}{3} \quad \Rightarrow \quad \kappa_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \mu_3$$

Also

$$\frac{\kappa_4}{4!} = \frac{\mu_4'}{4} - \frac{1}{2}\left(\frac{\mu_2'^2}{4} + \frac{2\,\mu_1'\,\mu_3'}{3!}\right) + \frac{1}{3}\,\frac{3\,\mu_1'^2\,\mu_2'}{2} - \frac{\mu_1'^4}{4}$$

$$\Rightarrow \kappa_4 = \mu_4' - 3\,\mu_2'^2 - 4\,\mu_1'\,\mu_3' + 12\,\mu_1'^2\,\mu_2' - 6\,\mu_1'^4$$

$$= (\mu_4' - 4\mu_3'\,\mu_1' + 6\mu_2'\,\mu_1'^2 - 3\mu_1'^4) - 3(\mu_2'^2 - 2\mu_2'\,\mu_1'^2 + \mu_1'^4)$$

$$= \mu_4 - 3(\mu_2' - \mu_1'^2)^2 = \mu_4 - 3\mu_2^2 = \mu_4 - 3\kappa_2^2 \qquad (\because \mu_2 = \kappa_2)$$

$$\Rightarrow \qquad \mu_4 = \kappa_4 + 3\,\kappa_2^2$$

Hence we have obtained :

$$\left.\begin{aligned} &\text{Mean} = \kappa_1 \\ &\mu_2 = \kappa_2 = \text{Variance} \\ &\mu_3 = \kappa_3 \\ &\mu_4 = \kappa_4 + 3\,\kappa_2^2 \end{aligned}\right\} \qquad \text{...(6·62}b\text{)}$$

Remarks. 1. These results are of fundamental importance and should be committed to memory.

2. If we differentiate both sides of (6·62*a*) w.r.t. t 'r' times and then put $t = 0$, we get

$$\kappa_r = \left[\frac{d^r}{dt^r} K_X(t)\right]_{t=0} \qquad \text{...(6·62}c\text{)}$$

6·11·1. Additive Property of Cumulants. *The rth cumulant of the sum of independent random variables is equal to the sum of the rth cumulants of the individual variables.* Symbolically,

$$\kappa_r(X_1 + X_2 + \ldots + X_n) = \kappa_r(X_1) + \kappa_r(X_2) + \ldots + \kappa_r(X_n), \qquad \text{...(6·63)}$$

where X_i ; $i = 1, 2, \ldots, n$ are independent random variables.

Proof. We have, since X_i 's are independent.

$$M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t)\,M_{X_2}(t)\ldots M_{X_n}(t)$$

Taking logarithm of both sides, we get

$$K_{X_1+X_2+\ldots+X_n}(t) = K_{X_1}(t) + K_{X_2}(t) + \ldots + K_{X_n}(t)$$

Differentiating both sides w.r.t. t 'r' times and putting $t = 0$, we get

$$\left[\frac{d^r}{dt^r} K_{X_1+X_2+\ldots+X_n}(t)\right]_{t=0} = \left[\frac{d^r}{dt^r} K_{X_1}(t)\right]_{t=0}$$

$$+ \left[\frac{d^r}{dt^r} K_{X_2}(t)\right]_{t=0} + \ldots + \left[\frac{d^r}{dt^r} K_{X_n}(t)\right]_{t=0}$$

$$\Rightarrow \qquad \kappa_r(X_1 + X_2 + \ldots + X_n) = \kappa_r(X_1) + \kappa_r(X_2) + \ldots + \kappa_r(X_n),$$

which establishes the result.

6·11·2. Effect of Change of Origin and Scale on Cumulants. If we take

$$U = \frac{X - a}{h}, \text{ then } M_U(t) = \exp(-at/h)\,M_X(t/h)$$

$$\therefore \quad K_U(t) = -\frac{at}{h} + K_X(t/h)$$

$$\kappa_1' + \kappa_1' t + \kappa_2' \frac{t^2}{2!} + \ldots + \kappa_r' \frac{t^r}{r!} + \ldots = -\frac{at}{h} + \kappa_1 (t/h) + \kappa_2 \frac{(t/h)^2}{2!} + \ldots + \kappa_r \frac{(t/h)^r}{r!} + \ldots$$

where κ_r' and κ_r are the rth cumulants of U and X respectively.

Comparing coefficients, we get

$$\kappa_1' = \frac{\kappa_1 - a}{h} \quad \text{and} \quad \kappa_r' = \frac{\kappa_r}{h^r} \; ; \; r = 2, 3, \ldots \qquad \ldots(6\cdot63a)$$

Thus we see that except the first cumulant, all cumulants are independent of change of origin. But the cumulants are not invariant of change of scale as the rth cumulant of U is $(1/h^r)$ times the rth cumulant of the distribiution of X.

Example 6·37. *Let the random variable X assume the value 'r' with the probability law :*

$$P(X = r) = q^{r-1} p \; ; \; r = 1, 2, 3, \ldots$$

Find the m.g.f. of X and hence its mean and variance.

[Calicut Univ. B.Sc., Oct. 1992]

Solution. $M_X(t) = E(e^{tX})$

$$= \sum_{r=1}^{\infty} e^{tr} \; q^{r-1} p = \frac{p}{q} \sum_{r=1}^{\infty} (qe^t)^r$$

$$= \frac{p}{q} qe^t \sum_{r=1}^{\infty} (qe^t)^{r-1} = pe^t \left[1 + qe^t + (qe^t)^2 + \ldots\right]$$

$$= \left(\frac{pe^t}{1 - qe^t}\right)$$

If dash (′) denotes the differentiation w.r.t. t, then we have

$$M_X'(t) = \frac{pe^t}{(1 - qe^t)^2}, \quad M_X''(t) = pe^t \frac{(1 + qe^t)}{(1 - qe^t)^3}$$

$$\therefore \quad \mu_1' \text{ (about origin)} = M_X'(0) = \frac{p}{(1-q)^2} = \frac{1}{p}$$

$$\mu_2' \text{ (about origin)} = M_X''(0) = \frac{p(1+q)}{(1-q)^3} = \frac{1+q}{p^2}.$$

Hence $\quad \text{mean} = \mu_1' \text{ (about origin)} = \dfrac{1}{p}$

and $\quad \text{variance} = \mu_2 = \mu_2' - \mu_1'^2 = \dfrac{1+q}{p^2} - \dfrac{1}{p^2} = \dfrac{q}{p^2}$

Example 6·38. *The probability density function of the random variable X follows the following probability law :*

$$p(x) = \frac{1}{2\theta}\, exp\left(-\frac{|x-\theta|}{\theta}\right), \; -\infty < x < \infty$$

Find M.G.F. of X. Hence or otherwise find E (X) and V (X).

[Punjab Univ. M.A.(Eco.), 1991]

Solution. The moment generating function of X is

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{2\theta} \exp\left(-\frac{|x-\theta|}{\theta}\right) e^{tx}\, dx$$

$$= \int_{-\infty}^{\theta} \frac{1}{2\theta} \exp\left(-\frac{|\theta-x|}{\theta}\right) e^{tx}\, dx$$

$$+ \int_{\theta}^{\infty} \frac{1}{2\theta} \exp\left(-\frac{|x-\theta|}{\theta}\right) e^{tx}\, dx$$

For $x \in (-\infty, \theta)$, $x - \theta < 0 \Rightarrow \theta - x > 0$

$\therefore \quad |x-\theta| = \theta - x \;\; \forall\, x \in (-\infty, \infty)$

Similarly, $\quad |x-\theta| = x - \theta \;\; \forall\, x \in (\theta, \infty)$

$$\therefore \quad M_X(t) = \frac{e^{-1}}{2\theta} \int_{-\infty}^{\theta} \exp\left[x\left(t + \frac{1}{\theta}\right)\right] dx + \frac{e}{2\theta} \int_{\theta}^{\infty} \exp\left[-x\left(\frac{1}{\theta} - t\right)\right] dx$$

$$= \frac{e^{-1}}{2\theta} \cdot \frac{1}{\left(t + \frac{1}{\theta}\right)} \cdot \exp\left[\theta\left(t + \frac{1}{\theta}\right)\right]$$

$$+ \frac{e}{2\theta} \cdot \frac{1}{\left(\frac{1}{\theta} - t\right)} \cdot \exp\left[-\theta\left(\frac{1}{\theta} - t\right)\right]$$

$$= \frac{e^{\theta t}.}{2(\theta t + 1)} + \frac{e^{\theta t}}{2(1 - \theta t)} = \frac{e^{\theta t}}{1 - \theta^2 t^2}$$

$$= e^{\theta t} (1 - \theta^2 t^2)^{-1}$$

$$= \left[1 + \theta t + \frac{\theta^2 t^2}{2!} + \ldots\right] \left[1 + \theta^2 t^2 + \theta^4 t^4 + \ldots\right]$$

$$= 1 + \theta t + \frac{3\,\theta^2 t^2}{2!} + \ldots$$

$\therefore \quad E(X) = \mu' = \text{Coefficient of } t \text{ in } M_X(t) = \theta$

$\mu_2' = \text{Coefficient of } \dfrac{t^2}{2!} \text{ in } M_X(t) = 3\,\theta^2$

Hence $\quad \text{Var}(X) = \mu_2' - \mu_1'^2 = 3\,\theta^2 - \theta^2 = 2\,\theta^2$

Example 6·39. *If the moments of a variate X are defined by*

$$E(X^r) = 0{\cdot}6\ ;\ r = 1, 2, 3, \ldots$$

show that $P(X = 0) = 0{\cdot}4, P(X = 1) = 0{\cdot}6, P(X \geq 2) = 0.$

[Delhi Univ. B.Sc. (Maths Hons.), 1985]

Solution. The m.g.f. of variate X is :

$$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r' = 1 + \sum_{r=1}^{\infty} \frac{t^r}{r!}(0{\cdot}6)$$

$$= 0{\cdot}4 + 0{\cdot}6 \sum_{r=0}^{\infty} \frac{t^r}{r!} = 0{\cdot}4 + 0{\cdot}6\, e^t \qquad \ldots(i)$$

But $\quad M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} P(X = x)$

$$= P(X = 0) + e^t . P(X = 1) + \sum_{x=2}^{\infty} e^{tx} . P(X = x) \qquad \ldots(ii)$$

From *(i)* and *(ii)*, we get :

$$P(X = 0) = 0{\cdot}4\ ;\ P(X = 1) = 0{\cdot}6\ ;\ P(X \geq 2) = 0.$$

Remark. In fact *(i)* is the m.g.f. of Bernoulli variate X with $P(X = 0) = q = 0{\cdot}4$ and $P(X = 1) = p = 0{\cdot}6$ [See § 7·1·2] and $P(X \geq 2) = 0$.

Example 6·40. *Find the moment generating function of the random variable whose moments are*

$$\mu_r' = (r + 1)!\, 2^r$$

Solution. The m.g.f. is given by

$$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r' = \sum_{r=0}^{\infty} \frac{t^r}{r!}(r + 1)!\, 2^r$$

$$= \sum_{r=0}^{\infty} (r + 1)(2t)^r$$

$$M_X(t) = 1 + 2 . (2t) + 3(2t)^2 + 4(2t)^3 + \ldots$$

$$= (1 - 2t)^{-2}$$

Aliter. The R.H.S. is an arithmetic-geometric series with ratio $r = (2t)$

Let $\quad S = 1 + 2r + 3r^2 + 4r^3 + \ldots$

Then $\quad rS = r + 2r^2 + 3r^3 + \ldots$

$\therefore \quad (1 - r)S = 1 + r + r^2 + \ldots = \dfrac{1}{(1 - r)}$

$\Rightarrow \quad S = \dfrac{1}{(1 - r)^2} = (1 - r)^{-2} = (1 - 2t)^{-2}$

Remark. This is the m.g.f. of Chi-square (χ^2) distribution with parameter (degrees of freedom) $n = 2$ [c.f. Chapter 13].

Example 6·41. *If μ_r' is the rth moment about origin, prove that*

$$\mu_r' = \sum_{j=1}^{r} \binom{r-1}{j-1} \mu_{r-j}' \kappa_j ,$$

where κ_j is the jth cumulant.

Solution. Differentiating both sides of (6·62*a*) in § 6·11, page 6·72 w.r.t. t, we get

$$\kappa_1 + \kappa_2 t + \kappa_3 \frac{t^2}{2\,!} + \ldots + \kappa_r \frac{t^{r-1}}{(r-1)\,!} + \ldots$$

$$= \frac{\mu_1' + \mu_2' t + \mu_3' \frac{t^2}{2\,!} + \ldots + \mu_r' \frac{t^{r-1}}{(r-1)\,!} + \ldots}{1 + \mu_1' t + \mu_2' \frac{t^2}{2\,!} + \ldots + \mu_r' \frac{t^r}{r\,!} + \ldots}$$

$$\Rightarrow \quad \left[\kappa_1 + \kappa_2 t + \kappa_3 \frac{t^2}{2\,!} + \ldots + \kappa_r \frac{t^{r-1}}{(r-1)\,!} + \ldots\right]$$

$$\times \left[1 + \mu_1' t + \mu_2' \frac{t^2}{2\,!} + \ldots + \mu_r' \frac{t^r}{r\,!} + \ldots\right]$$

$$= \mu_1' + \mu_2' t + \mu_3' \frac{t^2}{2\,!} + \ldots + \mu_r' \frac{t^{r-1}}{(r-1)\,!} + \ldots$$

Comparing the coefficient of $\frac{t^{r-1}}{(r-1)\,!}$ on both sides, we get

$$\mu_r' = \kappa_1 \cdot \mu_{r-1}' + (r-1)\,\kappa_2 \cdot \mu_{r-2}' + \binom{r-1}{2} \kappa_3 \cdot \mu_{r-3}' + \ldots + \kappa_r$$

$$= \binom{r-1}{0} \mu_{r-1}' \kappa_1 + \binom{r-1}{1} \mu_{r-2}' \kappa_2 + \binom{r-1}{2} \mu_{r-3}' \kappa_3 + \ldots + \binom{r-1}{r-1} \mu_0' \kappa_r$$

$$= \sum_{j=1}^{r} \binom{r-1}{j-1} \mu_{r-j}' \kappa_j ,$$

which is the required result.

6·12. Characteristic Function. In some cases m.g.f. does not exist, since the integral $\int_{-\infty}^{\infty} e^{tx} f(x)\, dx$ or the series $\sum_x e^{tx} p(x)$ does not converge absolutely for real values of t for some distributions. For example, for the continuous probability distribution

$$dF(x) = C \frac{1}{(1+x^2)^m}\, dx \;;\; m > 1, -\infty < x < \infty, \;.$$

the m.g.f. does not exist, since the integral

$$M_X(t) = C \int_{-\infty}^{\infty} e^{tx} \frac{1}{(1+x^2)^m}\, dx,$$

does not converge absolutely for finite positive values of m because the function e^{tx} dominates the function x^{2m} so that $e^{tx}/x^{2m} \to \infty$ as $x \to \infty$.

Again, for the discrete probability distribution

$$\left.\begin{aligned} f(x) &= \frac{6}{\pi^2 x^2}\ ;\ x = 1, 2, 3, \ldots \\ &= 0, \text{ elsewhere} \end{aligned}\right\}$$

$$M_X(t) = \sum_x e^{tx} f(x) = \frac{6}{\pi^2} \sum_{x=1}^{\infty} \left(\frac{e^{tx}}{x^2} \right)$$

The series is not convergent (by D'Alembert's Ratio Test) for $t > 0$. Thus there does not exist a positive number h such that $M_X(t)$ exists for $-h < t < h$. Hence $M_X(t)$ does not exist in this case also.

A more serviceable function than the m.g.f. is what is known as characteristic function and is defined as

$$\phi_X(t) = E(e^{itX}) = \int e^{itx} f(x)\, dx$$

(for continuous probability distributions)

$$= \sum_x e^{itx} f(x)$$

(for discrete probability distributions) ...(6·64)

If $F_X(x)$ is the distsribution function of a continuous random variable X, then

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx}\, dF(x) \qquad \text{...(6·64a)}$$

Obviously $\phi(t)$ is a complex valued function of real variable t. It may be noted that

$$|\phi(t)| = \left| \int e^{itx} f(x)\, dx \right| \le \int |e^{itx}|\, f(x)\, dx = \int f(x)\, dx = 1,$$

since $\quad |e^{itx}| = |\cos tx + i \sin tx|^{1/2} = (\cos^2 tx + \sin^2 tx)^{1/2} = 1$

Since $|\phi(t)| \le 1$, characteristic function $\phi_X(t)$ always exists.

Yet another advantage of characteristic function lies in the fact that it uniquely determines the distribution function, *i.e.*, if the characteristic function of a distribution is given, the distribution can be uniquely determined by the theorem, known as the Uniqueness Theorem of Characteristic Functions [c.f. Theorem 6·27 page 6·90].

6·12·1. Properties of Characteristic Functions. For all real 't', we have

(i) $\phi(0) = \int_{-\infty}^{\infty} dF(x) = 1$ ...(6·64b)

(ii) $|\phi(t)| \le 1 = \phi(0)$ ...(6·64c)

(iii) $\phi(t)$ is continuous everywhere, *i.e.*, $\phi(t)$ is a continuous function of t in $(-\infty, \infty)$. Rather $\phi(t)$ is uniformly continuous in 't'.

Proof. For $h \ne 0$, $|\phi_X(t+h) - \phi_X(t)| = \left| \int_{-\infty}^{\infty} \left[e^{i(t+h)x} - e^{itx} \right] dF(x) \right|$

$$\leq \int_{-\infty}^{\infty} \left| e^{itx} (e^{ihx} - 1) \right| dF(x)$$

$$= \int_{-\infty}^{\infty} \left| e^{ihx} - 1 \right| dF(x) \qquad \ldots(*)$$

The last integral does not depend on t. If it tends to zero as $h \to 0$ then $\phi_X(t)$ is uniformly continuous in 't'.

Now $\quad \left| e^{ihx} - 1 \right| \leq \left| e^{ihx} \right| + 1 = 2$

$$\therefore \quad \int_{-\infty}^{\infty} \left| e^{ihx} - 1 \right| dF(x) \leq 2 \int_{-\infty}^{\infty} dF(x) = 2.$$

Hence by Dominated Convergence Theorem (D.C.T.), taking the limit inside the integral sign in (*), we get

$$\lim_{h \to 0} \left| \phi_X(t+h) - \phi_X(t) \right| \leq \int_{-\infty}^{\infty} \lim_{h \to 0} \left| e^{ihx} - 1 \right| dF(x) = 0$$

$$\Rightarrow \quad \lim_{h \to 0} \phi_X(t+h) = \phi_X(t) \ \forall \ t.$$

Hence $\phi_X(t)$ is uniformly continuous in 't'.

(*iv*) $\phi_X(-t)$ and $\phi_X(t)$ are conjugate functions, *i.e.*, $\phi_X(-t) = \overline{\phi_X(t)}$, where $\bar{a}$ is the complex conjugate of a.

Proof. $\quad \phi_X(t) = E(e^{itX}) = E[\cos tX + i \sin tX]$

$$\Rightarrow \quad \overline{\phi_X(t)} = E[\cos tX - i \sin tX]$$
$$= E[\cos(-t)X + i \sin(-t)X]$$
$$= E(e^{-itX}) = \phi_X(-t).$$

6·12·2. Theorems on Characteristic Function.

Theorem 6·20. *If the distribution function of a r.v. X is symmetrical about zero, i.e.,*

$$1 - F(x) = F(-x) \quad \Rightarrow \quad f(-x) = f(x),$$

then $\phi_X(t)$ is real valued and even function of t.

Proof. By definition we have

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx = \int_{-\infty}^{\infty} e^{-ity} f(-y)\, dy \qquad (x = -y)$$

$$= \int_{-\infty}^{\infty} e^{-ity} f(y)\, dy \qquad [\because f(-y) = f(y)]$$

$$= \phi_X(-t) \qquad \ldots(*)$$

$\Rightarrow \quad \phi_X(t)$ is an even function of t.

Also $\quad \overline{\phi_X(t)} = \phi_X(-t) \qquad$ [c.f. Property (*iv*) § 6·12·1.]

$\therefore \quad \overline{\phi_X(t)} = \phi_X(-t) = \phi_X(t) \qquad$ (From *)

Hence $\phi_X(t)$ is a real valued function of t.

Theorem 6·21. *If X is some random variable with characteristic function $\phi_X(t)$, and if $\mu_r' = E(X^r)$ exists, then*

$$\mu_r' = (-i)^r \left[\frac{\partial^r}{\partial t^r} \phi(t) \right]_{t=0} \quad ...(6\cdot65)$$

Proof. $\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx$

Differentiating (under the integral sign) 'r' times w.r.t. t, we get

$$\frac{\partial^r}{\partial t^r} \phi(t) = \int_{-\infty}^{\infty} (ix)^r \cdot e^{itx} f(x)\, dx = (i)^r \int_{-\infty}^{\infty} x^r e^{itx} f(x)\, dx$$

$$\therefore \left[\frac{\partial^r}{\partial t^r} \phi(t) \right]_{t=0} = (i)^r \left[\int_{-\infty}^{\infty} x^r e^{itx} f(x)\, dx \right]_{t=0}$$

$$= (i)^r \int_{-\infty}^{\infty} x^r f(x)\, dx = i^r E(X^r) = i^r \mu_r'$$

Hence $\mu_r' = \left(\frac{1}{i} \right)^r \left[\frac{\partial^r}{\partial t^r} \phi(t) \right]_{t=0} = (-i)^r \left[\frac{\partial^r}{\partial t^r} \phi(t) \right]_{t=0}$

The theorems, *viz.*, 6·17, 6·18 and 6·19 on m.g.f. can be easily extended to the characteristic functions as given below.

Theorem 6·22. $\phi_{cX}(t) = \phi_X(ct)$, *c, being a constant.*

Theorem 6·23. *If X_1 and X_2 are independent random variables, then*

$$\phi_{X_1 + X_2}(t) = \phi_{X_1}(t)\, \phi_{X_2}(t) \quad ...(*)$$

More, generally for independent random variables X_i ; $i = 1, 2, ..., n$, we have

$$\phi_{X_1 + X_2 + ... + X_n}(t) = \phi_{X_1}(t)\, \phi_{X_2}(t) \ldots \phi_{X_n}(t)$$

Important Remark. Converse of (*) is not true, *i.e.*,

$\phi_{X_1 + X_2}(t) = \phi_{X_1}(t)\, \phi_{X_2}(t)$ does not imply that X_1 and X_2 are independent.

For example, let X_1 be a standard Cauchy variate with p.d.f.

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty$$

Then $\phi_{X_1}(t) = e^{-|t|}$ (c.f. Chapter 8)

Let $X_2 \equiv X_1$, *i.e.*, $P(X_1 = X_2) = 1$. ...(**)

Then $\phi_{X_2}(t) = e^{-|t|}$

Now $\phi_{X_1 + X_2}(t) = \phi_{2X_1}(t) = \phi_{X_1}(2t) = e^{-2|t|}$

$$= \phi_{X_1}(t)\, \phi_{X_2}(t)$$

i.e., (*) is satisfied but obviously X_1 and X_2 are not independent, because of (**).

In fact, (*) will hold even if we take $X_1 = aX$ and $X_2 = bX$, a and b being real numbers so that X_1 and X_2 are connected by the relation :

$$\frac{X_1}{a} = X = \frac{X_2}{b} \quad \Rightarrow \quad aX_2 = bX_1 .$$

As another example let us consider the joint p.d.f. of two random variables X and Y given by

$$f(x, y) = \frac{1}{4a^2} [1 + xy(x^2 - y^2)] \, ; \; |x| \le a, \; |y| \le a, \; a > 0$$

$$= 0, \text{ elsewhere}$$

Then the marginal p.d.f.'s of X and Y are given by

$$g(x)=\int_{-a}^{a} f(x,y)\,dy=\frac{1}{2a}\,;\ |x|\le a \qquad \text{(on simplification)}$$

$$h(y)=\int_{-a}^{a} f(x,y)\,dx=\frac{1}{2a}\,;\ |y|\le a$$

Then

$$\phi_X(t)=\int_{-\infty}^{\infty} e^{itx} g(x)\,dx=\frac{1}{2a}\int_{-a}^{a} e^{itx}\,dx$$

$$=\frac{e^{iat}-e^{-iat}}{2ait}=\frac{\sin at}{at}$$

Similarly $$\phi_Y(t)=\frac{\sin at}{at}$$

$$\therefore \qquad \phi_X(t)\ \phi_Y(t)=\left(\frac{\sin at}{at}\right)^2 \qquad \ldots(*)$$

The p.d.f. $k(z)$ of the random variable $Z=X+Y$ is given by the convolution of p.d.f.'s of X and Y, *viz.*,

$$k(z)=\int f(u,z-u)\,du$$

$$=\frac{1}{4a^2}\int \left[1+u(z-u)\{u^2-(z-u)^2\}\right]du$$

$$=\frac{1}{4a^2}\int (1+3z^2u^2-2zu^3-z^3u)\,du,$$

the limits of integration for u being in terms of z and are given by (left as an exercise to the reader)

$$-a\le u\le z+a\,;\ u\le 0$$

$$z-a\le u\le a\,;\ u>0$$

and

Thus

$$k(z)=\begin{cases}\dfrac{1}{4a^2}\displaystyle\int_{-a}^{z+a} (1+3z^2u^2-2zu^3-z^3u)\,du=\dfrac{2a+z}{4a^2}\,; & -2a\le z\le 0\\[2ex] \dfrac{1}{4a^2}\displaystyle\int_{z-a}^{a} (1+3z^2u^2-2zu^3-z^3u)\,du=\dfrac{2a-z}{4a^2}\,; & 0<z\le 2a\\[2ex] 0, \text{ elsewhere} \end{cases}$$

Now

$$\phi_{X+Y}(t)=\phi_Z(t)=\int_{-2a}^{2a} e^{itz}\,k(z)\,dz$$

$$=\int_{-2a}^{0}\left(\frac{2a+z}{4a^2}\right)e^{itz}\,dz+\int_{0}^{2a}\left(\frac{2a-z}{4a^2}\right)e^{itz}\,dz$$

$$= \int_0^{2a} \left(e^{-itz} + e^{itz}\right)\left(\frac{2a-z}{4a^2}\right). dz$$

[Changing z to $-z$ in the first integral]

$$= \frac{1}{2a^2}\int_0^{2a} (2a-z)\cos tz\, dz$$

$$= \frac{2-2\cos 2at}{4a^2t^2} = \frac{1-\cos 2at}{2a^2t^2} \quad \text{(on simplification)}$$

$$= \left(\frac{\sin at}{at}\right)$$

$$= \phi_X(t) . \phi_Y(t) \quad \text{[From (*)]}$$

But $\quad g(x) . h(y) \neq f(x, y)$

$\Rightarrow \quad X$ and Y are not independent.

However

$$\phi_{X_1, X_2}(t_1, t_2) = E\left(e^{it_1X_1 + it_2X_2}\right) = \phi_{X_1}(t) . \phi_{X_2}(t)$$

implies that X_1 and X_2 are independent.

(For proof see Theorem 6·28)

Theorem 6·24. *Effect of Change of Origin and Scale on Characteristic Function. If* $U = \frac{X-a}{h}$, *a and h being constants, then*

$$\phi_U(t) = e^{-iat/h}\, \phi_X(t/h)$$

In particular if we take $a = E(X) = \mu$ (say) and $h = \sigma_X = \sigma$ then the characteristic function of the standard variate

$$Z = \frac{X - E(X)}{\sigma_X} = \frac{X-\mu}{\sigma},$$

is given by $\quad \phi_Z(t) = e^{-i\mu t/\sigma}\, \phi_X(t/\sigma) \quad$ ...(6·66)

Definition. *A random variable X is said to be a Lattice variable or be lattice distributed, if for some* $h > 0$,

$$P\left[\frac{X}{h} \text{ is an integer}\right] = 1,$$

h is called a mesh.

Theorem 6·25. *If* $|\phi_X(s)| = 1$ *for some* $s \neq 0$, *then for some real a,* $X - a$ *is a Lattice variable with mesh* $h = 2\pi/|s|$.

Proof. Consider any fixed t. We can write

$\phi_X(t) = |\phi_X(t)|\, e^{iat}$, *(a dependent on t), since any complex number z can be written as* $z = |z|\, e^{i\theta}$.

$$\therefore \quad |\phi_X(t)| = e^{-iat}\phi_X(t) = \phi_{(X-a)}(t)$$

$$= E[\cos t(X-a) + i \sin t(X-a)] = E[\cos t(X-a)]$$

since left-hand side being real, we must have $E[\sin t(X-a)] = 0$.

$$\therefore \quad 1 - |\phi_X(t)| = E[1 - \cos t(X-a)] \quad \text{...(*)}$$

If $|\phi_X(s)| = 1, s \neq 0$ then for some a dependent on s, we have from (*)

$$E[1 - \cos s(X - a)] = 0 \qquad ...(**)$$

But since $1 - \cos s(X - a)$ is a non-negative random variable, (**)

$\Rightarrow \quad P[1 - \cos s(X - a) = 0] = 1$

$\Rightarrow \quad P[\cos s(X - a) = 1] = 1$

$\Rightarrow \quad P[s(X - a) = 2n\pi] = 1$

$\Rightarrow \quad P\left[(X - a) = \dfrac{2n\pi}{|s|}\right] = 1$, for some $n = 0, 1, 2, ...$

Thus $(X - a)$ is a Lattice variable with mesh $h = \dfrac{2\pi}{|s|}$.

6·12·3. Necessary and Sufficient Conditions for a Function $\phi(t)$ to be Characteristic Function. Properties *(i)* to *(iv)* in § 6·12·1 are merely the necessary conditions for a function $\phi(t)$ to be the characteristic function of an r.v. X. Thus if a function $\phi(t)$ does not satisfy any one of these four conditions, it cannot be the characteristic function of an r.v. X. For example, the function

$$\phi(t) = \log(1 + t),$$

cannot be the c.f. of r.v. X since $\phi(0) = \log 1 = 0 \neq 1$.

These conditions are, however, not sufficient. It has been shown (*c.f.* Methods of Mathematical Statistics by H. Cramer) that if $\phi(t)$ is near $t = 0$ of the form,

$$\phi(t) = 1 + 0(t^{2} + \delta), \delta > 0 \qquad ...(*)$$

where $0(t^r)$ divided by t^r tends to zero as $t \to 0$, then $\phi(t)$ cannot be the characteristic function unless it is identically equal to one. Thus, the functions

(i) $\phi(t) = e^{-t^4} = 1 + 0(t^4)$

(ii) $\phi(t) = \dfrac{1}{1 + t^4} = 1 + 0(t^4)$

being of the form (*) are not characteristic functions, though both satisfy all the necessary conditions.

We give below a set of sufficient but not necessary conditions, due to Polya for a function $\phi(t)$ to be the characteristic function :

$\phi(t)$ is a characteristic function if

(1) $\phi(0) = 1$,

(2) $\phi(t) = \phi(-t)$

(3) $\phi(t)$ is continuous

(4) $\phi(t)$ is convex for $t > 0$, *i.e.*, for $t_1, t_2 > 0$,

$$2\phi\left[\tfrac{1}{2}(t_1 + t_2)\right] \leq \phi(t_1) + \phi(t_2)$$

(5) $\lim\limits_{t \to \infty} \phi(t) = 0$.

Hence by Polya's conditions the functions $e^{-|t|}$ and $[1 + |t|]^{-1}$ are characteristic functions. However, Polya's conditions are only sufficient and not necessary for a characteristic function. For example, if $X \sim N(\mu, \sigma^2)$,

$$\phi(t) = e^{it\mu - t^2\sigma^2/2} \qquad [cf. \S\ 8{\cdot}5]$$

and $\qquad \phi(-t) \neq \phi(t)$.

Various necessary and sufficient conditions are known, the simplest seems to be the following, due to Cramer.

"In order that a given, bounded and continuous function $\phi(t)$ should be the characteristic function of a distribution, it is necessary and sufficient that $\phi(0) = 1$ and that the function

$$\phi(x, A) = \int_0^A \int_0^A \phi(t-u)\, e^{(t-u)ix} dt\, du$$

is real and non-negative for all real x and all $A > 0$.

6·12·4. Multi-variate Characteristic Function. Then

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \text{ and } t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}, \; t \textit{ real}$$

be $z \times 1$ column vectors. Then characteristic function of X is defined as

$$\phi_X(t) = E(e^{it'X}) = E[e^{i(t_1X_1 + t_2X_2 + \dots + t_nX_n)}] \qquad \dots(6{\cdot}67)$$

We may also write it as

$$\phi_{X_1, X_2, \dots X_n}(t_1, t_2, \dots, t_n) \text{ or } \phi_X(t_1, t_2, \dots, t_n)$$

Some Properties.

(i) $\phi_X(0, 0, \dots 0) = 1$

(ii) $\phi_{-X}(t) = \overline{\phi_X(t)}$

(iii) $|\phi_X(t)| \le 1$

(iv) $\phi_X(t)$ is uniformly continuous in n-dimensional Euclidian space.

(v) If $f_X(\cdot)$ is p.d.f. of X,

$$\phi_X(t) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_X(x_1, x_2, \dots, x_n)\, e^{i\Sigma t_j x_j}\, dx_1\, dx_2 \dots dx_n$$

(vi) $\phi_X(t) = \phi_Y(t)$ for all t, then X and Y have the same distribuiton.

(vii) If $\displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\phi_X(t_1, t_2, \dots t_n)|\, dt_1\, dt_2 \dots dt_n < \infty$,

then X is absolutely continuous and has a uniformly continuous p.d.f.

$$f_X(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i\Sigma t_j x_j}\, \phi_X(t_1, t_2, \dots, t_n)\, dt_1\, dt_2 \dots dt_n$$

(viii) The random variables $X_1, X_2, \dots, X_n$ are (mutually) independent iff

$$\phi_{X_1, X_2, \dots X_n}(t_1, t_2, \dots, t_n) = \phi_{X_1}(t_1)\, \phi_{X_2}(t_2) \dots \phi_{X_n}(t_n)$$

Remark. *Multivariate Moment Generating Function.* Similarly, the m.g.f. of vector $X = (X_1, X_2, \dots, X_n)'$ is given by :

$$M_X(t) = E(e^{t'X}) = E(e^{t_1X_1 + t_2X_2 + \dots t_nX_n}) \qquad \dots(6\cdot68)$$

We may also write :

$$M_X(t) = M_{X_1, X_2, \dots X_n}(t_1, t_2, \dots, t_n) = E(e^{t_1X_1 + t_2X_2 + \dots t_nX_n})$$

In particular, for two variates X_1 and X_2

$$M_X(t) = M_{X_1, X_2}(t_1, t_2) = E(e^{t_1X_1 + t_2X_2}) = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \frac{t_1^{\ r}}{r!} \frac{t_2^{\ s}}{s!} E(X_1^{\ r} X_2^{\ s}), \qquad \dots(6\cdot69)$$

provided it exists for $-h_1 < t_1 < h_1$ and $-h_2 < t_2 < h_2$, where h_1 and h_2 are positive.

$$M_{X_1, X_2}(t_1, 0) = E(e^{t_1X_1}) = M_{X_1}(t_1) \qquad \dots(6\cdot69a)$$

$$M_{X_1, X_2}(0, t_2) = E(e^{t_2X_2}) = M_{X_2}(t_2) \qquad \dots(6\cdot69b)$$

If $M(t_1, t_2)$ exists, the moments of all orders of X and Y exist and are given by:

$$E(X_2^{\ r}) = \left[\frac{\partial^r M(t_1, t_2)}{\partial t_2^{\ r}}\right]_{t_1 = t_2 = 0} = \frac{\partial^r M(0, 0)}{\partial t_2^{\ r}} \qquad \dots(6\cdot70)$$

$$E(X_1^{\ r}) = \left[\frac{\partial^r M(t_1, t_2)}{\partial t_1^{\ r}}\right]_{t_1 = t_2 = 0} = \frac{\partial^r M(0, 0)}{\partial t_1^{\ r}} \qquad \dots(6\cdot70a)$$

$$E(X_1^{\ r} X_2^{\ s}) = \left[\frac{\partial^{r+s} M(t_1, t_2)}{\partial t_1^{\ r} \partial t_2^{\ s}}\right]_{t_1 = t_2 = 0} = \frac{\partial^{r+s} M(0, 0)}{\partial t_1^{\ r} \partial t_2^{\ s}} \qquad \dots(6\cdot70b)$$

Cumulant generating function of $X = (X_1, X_2)'$ is given by :

$$K_{X_1, X_2}(t_1, t_2) = \log M_{X_1, X_2}(t_1, t_2). \qquad \dots(6\cdot71)$$

Example 6·42. *For a distribution, the cumulants are given by*

$$\kappa_r = n[(r-1)!], \quad n > 0$$

Find the characteristic function. **(Delhi Univ. B.Sc. (Stat. Hons.), 1990)**

Solution. The cumulant generating function $K(t)$, if it exists, is given by

$$K(t) = \sum_{r=1}^{\infty} \frac{(it)^r}{r!} \kappa_r = \sum_{r=1}^{\infty} \frac{(it)^r}{r!} n\{(r-1)!\} = n \sum_{r=1}^{\infty} \frac{(it)^r}{r}$$

$$= n\left[it + \frac{(it)^2}{2} + \frac{(it)^3}{3} + \dots\right] = n[\log(1 - it)]$$

$$= -n \log(1 - it) = \log(1 - it)^{-n}$$

Also we have

$$K(t) = \log \phi(t) = \log(1 - it)^{-n}$$

$$\therefore \qquad \phi(t) = (1 - it)^{-n}$$

Remark. This is the characteristic function of the gamma distribution : (*c.f.* § 8·3·1)

$$f(x) = \frac{e^{-x}\, x^{n-1}}{\Gamma(n)}\ ;\ n > 0,\ 0 < x < \infty.$$

Example 6·43. *The moments about origin of a distribution are given by*

$$\mu_r' = \frac{\Gamma(\nu + r)}{\Gamma(\nu)}.$$

Find the characteristic function.

(Madurai Kamaraj Univ. B.Sc., 1990)

Solution. We have

$$\phi(t) = \sum_{r=0}^{\infty} \frac{(it)^r}{r\,!}\,\mu_r' = \sum_{r=0}^{\infty} \frac{(it)^r}{r\,!} \cdot \frac{\Gamma(\nu + r)}{\Gamma(\nu)}$$

$$= \sum_{r=0}^{\infty} \frac{(it)^r}{r\,!} \cdot \frac{(\nu + r - 1)\,!}{(\nu - 1)\,!}$$

$$= \sum_{r=0}^{\infty} (it)^r \cdot {}^{\nu + r - 1}C_r = \sum_{r=0}^{\infty} (-1)^r \cdot {}^{-\nu}C_r\, (it)^r$$

$$[\because\ {}^{-\nu}C_r = (-1)^r \cdot {}^{\nu + r - 1}C_r \Rightarrow (-1)^r \cdot {}^{-\nu}C_r = {}^{\nu + r - 1}C_r]$$

$$\therefore \quad \phi(t) = \sum_{r=0}^{\infty} {}^{-\nu}C_r\, (-it)^r = (1 - it)^{-\nu}$$

Example 6·44. *Show that*

$$e^{itx} = 1 + (e^{it} - 1)\, x^{(1)} + (e^{it} - 1)^2 \frac{x^{(2)}}{2\,!} + \ldots + (e^{it} - 1)^r \cdot \frac{x^{(r)}}{r\,!} + \ldots$$

where $x^{(r)} = x(x-1)(x-2)\ldots(x-r+1)$. *Hence show that*

$\mu_{(r)}' = [D^r\, \phi(t)]_{t=0}$, *where* $D = \dfrac{d}{d(e^{it})}$ *and* $\mu_{(r)}'$ *is the* r^{th} *factorial moment.*

Solution. We have

$$\text{R.H.S.} = 1 + (e^{it} - 1)\, x^{(1)} + (e^{it} - 1)^2 \cdot \frac{x^{(2)}}{2\,!} + \ldots + (e^{it} - 1)^r \cdot \frac{x^{(r)}}{r\,!} + \ldots$$

$$= 1 + (e^{it} - 1)({}^xC_1) + (e^{it} - 1)^2 ({}^xC_2)$$
$$+ (e^{it} - 1)^3 ({}^xC_3) + \ldots + (e^{it} - 1)^r ({}^xC_r)$$

$$= [1 + (e^{it} - 1)]^x = e^{itx} = \text{L.H.S.}$$

By def.

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx$$

$$= \int_{-\infty}^{\infty} \left[1 + (e^{it} - 1)x^{(1)} + (e^{it} - 1)^2 \cdot \frac{x^{(2)}}{2!} + \cdots + (e^{it} - 1)^r \cdot \frac{x^{(r)}}{r!} + \cdots \right] f(x)\, dx$$

$$= 1 + (e^{it} - 1) \int_{-\infty}^{\infty} x^{(1)} f(x)\, dx + \frac{(e^{it} - 1)^2}{2!} \int_{-\infty}^{\infty} x^{(2)} f(x)\, dx + \cdots$$

$$+ \frac{(e^{it} - 1)^r}{r!} \int_{-\infty}^{\infty} x^{(r)} f(x)\, dx + \cdots$$

$$\therefore \quad [D^r \phi(t)]_{t=0} = \left[\frac{d^r \phi(t)}{d(e^{it})^r} \right]_{t=0} = \int_{-\infty}^{\infty} x^{(r)} f(x)\, dx = \mu_{(r)}'$$

where $\mu_{(r)}'$ is the rth factorial moment.

Theorem 6·26. (Inversion Theorem). **Lemma.** *If $(a-h, a+h)$ is the continuity interval of the distribution function $F(x)$, then*

$$F(a+h) - F(a-h) = \lim_{T \to \infty} \frac{1}{\pi} \int_{-T}^{T} \frac{\sin ht}{t} e^{-ita} \phi(t)\, dt,$$

$\phi(t)$ being the characteristic function of the distribution.

Corollary. If $\phi(t)$ is absolutely integrable over R^1, *i.e.*, if

$$\int_{-\infty}^{\infty} |\phi(t)|\, dt < \infty,$$

then the derivative of $F(x)$ exists, which is bounded, continuous on R^1 and is given by

$$f(x) = F'(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t)\, dt, \qquad \text{...(6·72)}$$

for every $x \in R^1$.

Proof. In the above lemma replacing a by x and on dividing by $2h$, we have

$$\frac{F(x+h) - F(x-h)}{2h} = \frac{1}{2\pi} \cdot \lim_{T \to \infty} \int_{-T}^{T} \frac{\sin ht}{ht} e^{itx} \phi(t)\, dt$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sin ht}{ht} e^{-itx} \phi(t)\, dt$$

$$\therefore \quad \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} = \frac{1}{2\pi} \lim_{h \to 0} \int_{-\infty}^{\infty} \frac{\sin ht}{ht} \cdot e^{-itx} \phi(t)\, dt$$

Since $\qquad \int_{-\infty}^{\infty} |\phi(t)|\, dt < \infty,$

the integrand on the right hand side is bounded by an integrable function and hence by Dominated Convergence Theorem, we get

$$\lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \lim_{h \to 0} \left(\frac{\sin ht}{ht} \right) \cdot e^{-itx} \phi(t)\, dt$$

By mean value theorem of differential calculus, we have

$$\lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} = F'(x) = f(x),$$

where $f(.)$ is the p.d.f. corresponding to $\phi(t)$. Thus

$$f(x) = F'(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)\,dt,$$

as desired.

Remark. Consider the function $\mathcal{F}_c$ defined by

$$\mathcal{F}_c = \int_{-c}^{c} e^{-itx}\,\phi(t)\,dt$$

Now if $F'(x) = f(x)$ exists, then

$$\lim_{c\to\infty}\frac{\mathcal{F}_c}{2c} = \lim_{c\to\infty}\frac{1}{2c}\int_{-c}^{c}\phi(t)\,e^{-itx}\,dt$$

$$= \lim_{c\to\infty}\left\{\frac{1}{2c}\,.\,2\pi f(x)\right\} = 0$$

Hence $\dfrac{\mathcal{F}_c}{2c}\to 0$ at all points where $F(x)$ is continuous. In other words, if the prbability distribution is continuous

$$\frac{\mathcal{F}_c}{2c}\to 0 \text{ as } c\to\infty$$

If, however, the frequency function is discontinuous, *i.e.*, distribution is discrete, consider one point of discontinuity say, the frequency f_j at $x = x_j$. Then the contribution of x_j to $\phi(t)$ is $f_j e^{itx_j}$ and hence its contributions to $\mathcal{F}_c$ will be

$$\int_{-c}^{c} f_j e^{itx_j} e^{-itx}\,dt$$

$$\therefore\quad \lim_{c\to\infty}\frac{\mathcal{F}_c}{2c} = \lim_{c\to\infty}\frac{1}{2c} f_j\int_{-c}^{c} e^{it(x_j - x)}\,dt$$

$$= \lim_{c\to\infty}\frac{1}{2c} f_j\left[\frac{e^{it(x_j-x)}}{i(x_j-x)}\right]_{-c}^{c}$$

$$= \begin{cases} 0 & \text{for } x \neq x_j \\ f_j & \text{for } x = x_j \end{cases}$$

Hence if $\mathcal{F}_c/2c\to 0$ at a point, there is no discontinuity in the distribution function at that point, but if it tends to a positive number f_j, the distribution is discontinuous at that point and the frequency is f_j. This gives us a criterion whether a given characteristic function represents a continuous distribution or not.

Theorem 6·27. Uniqueness Theorem of Characteristic Functions. *Characteristic function uniquely determines the distribution, i.e., a necessary and sufficient condition for two distributions with p.d.f.'s $f_1(.)$ and $f_2(.)$ to be identical is that their characteristic functions $\phi_1(t)$ and $\phi_2(t)$ are identical.*

Proof. If $f_1(\cdot)=f_2(\cdot)$, then from the definition of characteristic function, we get

$$\phi_1(t)=\int_{-\infty}^{\infty} e^{itx} f_1(x)\,dx=\int_{-\infty}^{\infty} e^{itx} f_2(x)\,dx=\phi_2(t)$$

Conversely if $\phi_1(t)=\phi_2(t)$, then from corollary to Theorem 6·26, we get

$$f_1(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,\phi_1(t)\,dt=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,\phi_2(t)\,dt=f_2(x)$$

Remark. This is one of the most fundamental theorems in the distribution theory. It implies that corresponding to a distribution there is only one characteristic function and corresponding to a given characteristic function, there is only one distribution. This *one to one correspondence between characteristic functions and the p.d.f.'s enables us to identify the form of the p.d.f. from that of characteristic function.*

Theorem 6·28. *Necessary and sufficient condition for the random variables X_1 and X_2 to be independent is that their joint characteristic function is equal to the product of their individual characteristic functions, i.e.,*

$$\phi_{X_1,X_2}(t_1,t_2)=\phi_{X_1}(t_1)\,\phi_{X_2}(t_2) \qquad \ldots(*)$$

Proof. (*i*) *Condition is Necessary.* If X_1 and X_2 are independent then we have to show that (*) holds. By def.,

$$\begin{aligned}\phi_{X_1X_2}(t_2,t_2) &= E\left(e^{it_1X_1+it_2X_2}\right)=E\left(e^{it_1X_1}\cdot e^{it_2X_2}\right)\\ &= E\left(e^{it_1X_1}\right)E\left(e^{it_2X_2}\right)\ (\because X_1, X_2 \text{ are independent})\\ &= \phi_{X_1}(t_1)\,\phi_{X_2}(t_2),\end{aligned}$$

as required.

(*ii*) *Condition is sufficient.* We have to show that if (*) holds, then X_1 and X_2 are independent.

Let $f_{X_1,X_2}(x_1,x_2)$ be the joint p.d.f. of X_1 and X_2 and $f_1(x_1)$ and $f_2(x_2)$ be the marginal p.d.f.'s of X_1 and X_2 respectively. Then by definition (for continuous r.v.'s), we get

$$\phi_{X_1}(t_1)=\int_{-\infty}^{\infty} e^{it_1x_1} f_1(x_1)\,dx_1$$

$$\phi_{X_2}(t_2)=\int_{-\infty}^{\infty} e^{it_2x_2} f_2(x_2)\,dx_2$$

$$\begin{aligned}\therefore\quad \phi_{X_1}(t_1)\,\phi_{X_2}(t_2) &= \left[\int_{-\infty}^{\infty} e^{it_1x_1} f_1(x_1)\,dx_1\right]\left[\int_{-\infty}^{\infty} e^{it_2x_2} f_2(x_2)\,dx_2\right]\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(t_1x_1+t_2x_2)} f_1(x_1)\,f_2(x_2)\,dx_1\,dx_2 \qquad \ldots(**)\end{aligned}$$

by Fubini's theorem, since the integrand is bounded by an integrable function.

Also by def:
$$\phi_{X_1, X_2}(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i(t_1 x_1 + t_2 x_2)} f(x_1, x_2)\, dx_1\, dx_2$$

If (*) holds, we get from (**)
$$\phi_{X_1, X_2}(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i(t_1 x_1 + t_2 x_2)} f_1(x_1) f_2(x_2)\, dx_1\, dx_2$$

Hence by uniqueness theorem of characteristic functions, we get
$$f(x_1, x_2) = f_1(x_1) f_2(x_2),$$
which implies that X_1 and X_2 are independent.

Remarks. 1. For discrete r.v.'s, the result is established by using summation instead of integration.

2. The result can be generalised to the case of more than two variables.

Necessary and sufficient condition for the mutual independence of random variables X_i, $(i = 1, 2, 3, \ldots, n)$ is that
$$\phi_{X_1, X_2, \ldots, X_n}(t_1, t_2, \ldots, t_n) = \phi_{X_1}(t_1)\ \phi_{X_2}(t_2) \ldots \phi_{X_n}(t_n)$$

3. In terms of moment generating functions, the necessary and sufficient condition for the r.v.'s $X_1, X_2, \ldots, X_n$ to be mutually independent is that
$$M_{X_1, X_2, \ldots, X_n}(t_1, t_2, \ldots, t_n) = M_{X_1}(t_1)\ M_{X_2}(t_2) \ldots M_{X_n}(t_n)$$
provided m.g.f.'s exist.

Theorem 6·29. Hally–Bray Theorem. *If the sequence of distribution functions $\{F_n(x)\}$ converges to the distribution function $F(x)$ at all the points of continuity of the latter and $g(x)$ is bounded continuous function over the line $R^1(-\infty, \infty)$, then*

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} g(x)\, dF_n(x) = \int_{-\infty}^{\infty} g(x)\, dF(x) \qquad \ldots(*)$$

Corollary. If $F_n(x) \to F(x)$, then the corresponding sequence of characteristic functions $\phi_n(t)$ of $F_n(x)$ converges to the characteristic function $\phi(t)$ of F at every point 't'.

Proof. $\cos tx$ and $\sin tx$ are continuous and bounded functions of x for all t and hence from (*), we get

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} \cos tx\ dF_n(x) = \int_{-\infty}^{\infty} \cos tx\ dF(x)$$

and
$$\lim_{n \to \infty} \int_{-\infty}^{\infty} \sin tx\ dF_n(x) = \int_{-\infty}^{\infty} \sin tx\ dF(x)$$

$$\therefore \lim_{n \to \infty} \int_{-\infty}^{\infty} (\cos tx + i \sin tx)\ dF_n(x) = \int_{-\infty}^{\infty} (\cos tx + i \sin tx).$$

$$\Rightarrow \quad \lim_{n \to \infty} \int_{-\infty}^{\infty} e^{itx} \, dF_n(x) = \int_{-\infty}^{\infty} e^{itx} \, dF(x).$$

$$\Rightarrow \quad \phi_n(t) \to \phi(t) \text{ as } n \to \infty$$

Theorem 6·30. Continuity Theorem for Characteristic Functions. For a sequence of distribution functions $\{F_n(x)\}$ with the corresponding sequence of characteristic functions $\{\phi_n(t)\}$, a necessary and sufficient condition that $F_n(x) \to F(x)$ at all points of continuity of F is that for every real t, $\phi_n(t) \to \phi(t)$, which is continuous at $t = 0$ and $\phi(t)$ is the characteristic function corresponding to F.

Example 6·45. *Let $F_n(x)$ be the distribution function defined by*

$$\begin{aligned} F_n(x) &= 0 \text{ for } x \leq -n \\ &= \frac{x+n}{2n} \text{ for } -n < x < n \\ &= 1 \text{ for } x \geq n \end{aligned}$$

Is the limit $F_n(x)$ a distribution function ? If not, why?

Solution. $\phi_n(t) = \int_{-n}^{n} e^{itx} \frac{1}{2n} dx$ $\qquad (\because f_n(x) = F_n'(x))$

$$= \frac{1}{2n} \left[\frac{e^{itn} - e^{-itn}}{it} \right] = \frac{\sin nt}{nt}$$

$$\phi(t) = \lim_{n \to \infty} \phi_n(t) = \lim_{n \to \infty} \frac{\sin nt}{nt}$$

$$= \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases}$$

i.e., $\phi(t)$ is discontinuous at $t = 0$.

and $\qquad \lim_{n \to \infty} F_n(x) = \frac{1}{2}$

Hence $F(x)$ is not a distribution function.

Example 6·46. *Find the distribution for which characteristic function is* (*a*) $\phi(t) = (q + pe^{it})^n$, (*b*) $\phi(t) = e^{-t^2\sigma^2/2}$

Solution. (*a*) $\qquad \phi(t) = (q + pe^{it})^n = \sum_{j=0}^{n} {}^nC_j p^j q^{n-j} e^{itj}$

We have

$$\mathcal{F}_c = \int_{-c}^{c} e^{-itx} \phi(t) \, dt = \int_{-c}^{c} \left\{ e^{-itx} \sum_{j=0}^{n} {}^nC_j \, p^j q^{n-j} e^{itj} \right\} dt$$

$$= \sum_{j=0}^{n} \left[{}^nC_j p^j q^{n-j} \int_{-c}^{c} e^{-it(x-j)} dt \right]$$

(*i*) If $x \neq j$,

$$\mathcal{F}_c = \sum_{j=0}^{n} {}^nC_j p^j q^{n-j} \left| \frac{e^{-it(x-j)}}{-i(x-j)} \right|_{-c}^{c} = \sum_{j=0}^{n} {}^nC_j p^j q^{n-j} \left[\frac{e^{ic(x-j)} - e^{-ic(x-j)}}{i(x-j)} \right]$$

$$= \sum_{j=0}^{n} \left[{}^nC_j p^j q^{n-j} \cdot \frac{2i \sin\{c(x-j)\}}{(x-j)} \right]$$

$$\therefore \qquad \lim_{c \to \infty} \frac{\mathcal{F}_c}{2c} \to 0 \ \forall x.$$

Hence there is no discontinuity in the distribution function when $x \neq j$.

(*ii*) If $x = j$,

$$\mathcal{F}_c = \sum_{j=0}^{n} \left[{}^nC_j p^j q^{n-j} \int_{-c}^{c} dt \right] = 2c \sum_{j=0}^{n} {}^nC_j p^j q^{n-j} = 2c\,(q+p)^n = 2c$$

Since $\dfrac{\mathcal{F}_c}{2c} \to 1$ at $x = j$, the distribution function is discontinuous and its frequency is ${}^nC_j p^j q^{n-j}$.

(*b*) Let $\displaystyle \mathcal{F}_c = \int_{-c}^{c} e^{-itx - \frac{1}{2}\sigma^2 t^2} dt$

$$\therefore \qquad |\mathcal{F}_c| \leq \int_{-c}^{c} \left| e^{-itx - \frac{1}{2}\sigma^2 t^2} \right| dt \leq \int_{-c}^{c} e^{-\frac{1}{2}\sigma^2 t^2} dt$$

$$\leq \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sigma^2 t^2} dt = \frac{\sqrt{2\pi}}{\sigma}$$

$$\therefore \qquad \lim_{c \to \infty} \frac{\mathcal{F}_c}{2c} = 0 \ \forall x.$$

Hence the distribution function is continuous for all x.

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} e^{-t^2 \sigma^2/2} dt$$

$$= \frac{1}{2\pi} e^{-x^2/2\sigma^2} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2} \left(t\sigma + \frac{ix}{\sigma} \right)^2 \right\} dt$$

Put $\qquad t\sigma + \dfrac{ix}{\sigma} = \xi$, *i.e.*, $\sigma dt = d\xi$

$$\therefore f(x)=\frac{1}{2\pi}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\int_{-\infty}^{\infty} e^{-\xi^2/2}\,\frac{d\xi}{\sigma}$$

Hence

$$f(x)=\frac{1}{2\pi}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\frac{\sqrt{2\pi}}{\sigma}=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\},\ -\infty<x<\infty$$

which is the p.d.f. of normal distribution.

Example 6·47. *Find the density function $f(x)$ corresponding to the characteristic function defined as follows :*

$$\phi(t)=\begin{cases}1-|t|, & |t|\le 1\\ 0, & |t|>1\end{cases}$$

[Delhi Univ, B.Sc. (Maths Hons.), 1989]

Solution. By Inversion Theorem, the p.d.f. of X is given by :

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,\phi(t)\,dt$$

$$=\frac{1}{2\pi}\int_{-1}^{0} e^{-itx}(1+t)\,dt+\frac{1}{2\pi}\int_{0}^{1} e^{-itx}(1-t)\,dt\ *$$

Now

$$\int_{-1}^{0} e^{-itx}(1+t)\,dt=\left[\frac{e^{-itx}}{-ix}(1+t)\right]_{-1}^{0}+\frac{1}{ix}\int_{-1}^{0} e^{-itx}\,dt$$

$$=-\frac{1}{ix}+\frac{1}{ix}\left[\frac{e^{-itx}}{-ix}\right]_{-1}^{0}$$

$$=-\frac{1}{ix}+\frac{1}{(ix)^2}(e^{ix}-1)$$

Similarly,

$$\int_{0}^{1} e^{-itx}(1-t)\,dt=\frac{1}{ix}+\frac{1}{(ix)^2}(e^{-ix}-1)$$

$$\therefore\qquad f(x)=\frac{1}{2\pi}\left[\frac{1}{(ix)^2}\left\{e^{ix}-1+e^{-ix}-1\right\}\right]$$

$$=\frac{1}{\pi x^2}\left[1-\frac{e^{ix}+e^{-ix}}{2}\right]=\frac{1}{\pi}\cdot\frac{1-\cos x}{x^2},\ -\infty<x<\infty$$

EXERCISE 6 (c)

1. Define m.g.f. of a random variable. Hence or otherwise find the m.g.f. of:

---

* $\because$ for $-1<t<0$, $|t|=-t$ and for $0<t<1$, $|t|=+t$

$$(i)\ Y = aX + b, \quad (ii)\ Y = \frac{X - m}{\sigma}.$$

[Sri Venkat Univ. B.Sc., Sept. 1990; Kerala Univ. B.Sc., Sept. 1992]

2. The random variable X takes the value n with probability $1/2^n$, $n = 1, 2, 3, \ldots$ Find the moment generating function of X and hence find the mean and variance of X.

3. Show that if $\overline{X}$ is mean of n independent random variables, then

$$M_{\overline{X}}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n$$

4. (*a*) Define moments and moment generating function (m.g.f.) of a random variable X. If $M(t)$ is the m.g.f. of a random variable X about the origin, show that the moment μ_r' is given by

$$\mu_r' = \left[\frac{d^r M(t)}{dt^r}\right]_{t=0}$$ **[Baroda Univ. B.Sc., 1992]**

(*b*) If μ_r' is the rth order moment about the origin and κ_j is the cumulant of jth order, prove that

$$\frac{\delta \mu_r'}{\delta \kappa_j} = \binom{r-1}{j-1} \mu'_{r-j}$$

(*c*) If μ_r' is the rth moment about the origin of a variable X and if $\mu_r' = r!$, find the m.g.f. of X.

5. (*a*) A random variable 'X' has probability function

$$p(x) = \frac{1}{2^x}\ ;\ x = 1, 2, 3, \ldots$$

Find the M.G.F., mean and variance.

(*b*) Show that the m.g.f. of r.v. X having the p.d.f.

$$f(x) = \frac{1}{3},\ -1 < x < 2$$
$$= 0, \text{ elsewhere,}$$

is
$$M(t) = \frac{e^{2t} - e^{-t}}{3t},\ t \neq 0$$
$$= 1,\ t = 0$$ **[Gujarat Univ. B.Sc., Oct. 1991]**

(*c*) A random variable 'X' has the density function :

$$f(x) = \frac{1}{2\sqrt{x}},\ 0 < x < 1$$
$$= 0, \text{ elsewhere}$$

Obtain the moment generating function and hence the mean and variance.

6. X is a random variable and $p(x) = ab^x$, where a and b are positive, $a + b = 1$ and x taking the values 0, 1, 2, ... Find the moment generating function of X. Hence show that

$$m_2 = m_1(2m_1 + 1)$$

m_1 and m_2 being the first two moments.

7. Find the characteristic function of the following distributions and their variances :

(i) $dF(x) = ae^{-ax}\,dx,\ (a > 0, x > 0)$

(ii) $dF(x) = \frac{1}{2}e^{-|x|}\,dx,\ (-\infty < x < \infty)$

(iii) $P(X = j)$ lineu$=\binom{n}{j}p^j q^{n-j},\ (0 < p < 1, q = 1 - p, j = 0, 1, 2, ..., n).$

8. Obtain the m.g.f. of the random variable X having p.d.f.,

$$f(x) = \begin{cases} x, & \text{for } 0 \le x < 1 \\ 2 - x, & \text{for } 1 \le x < 2 \\ 0, & \text{elsewhere} \end{cases}$$

Determine μ_1', μ_2, μ_3 and μ_4.

Ans. $\left(\dfrac{e^t - 1}{t}\right)^2$, $\mu_1' = 1$, $\mu_2' = 7$.

9. (a) Define cumulants and obtain the first four cumulants in terms of central moments.

(b) If X is a variable with zero mean and cumulants κ_r, show that the first two cumulants l_1 and l_2 of X^2 are given by $l_1 = \kappa_2$ and $l_2 = 2\kappa_2^2 + \kappa_4$.

10. Show that the rth cumulant for the distribution

$$f(x) = ce^{-cx}, \text{ where } c \text{ is positive and } 0 \le x < \infty$$

is $$\frac{1}{c^r}.(r-1)!$$

11. If X is a random variable with cumulants κ_r ; $r = 1, 2, ...$ Find the cumulants of

(i) cX, (ii) $c + X$, where c is a constant.

12. (a) Define the characteristic function of a random variable. Show that the characteristic function of the sum of two independent variables is equal to the product of their characteristic functions.

(b) If X is a random variable having cumulants κ_r ; $r = 1, 2, ...$ given by

$$\kappa_r = (r-1)!\,pa^{-r} ; p > 0, a > 0,$$

find the characteristic function of X.

(c) Prove that the characteristic function of a random variable X is real if and only if X has a symmetric distribution about 0.

13. Define $\phi(t)$, the characteristic function of a random variable. Find the characteristic function of a random variable X defined as follows :

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & 0 \le x \le 1 \\ 0, & x > 1 \end{cases}$$

Ans. $e^{(it-1)}/it$

14. For the joint distribution of two-dimensional random variable (X, Y) given by

$$f(x, y) = \frac{1}{4a^2}\left[1 + xy\,(x^2 - y^2)\right], \; |x| \le a\;;\; |y| \le a,\, a > 0$$
$$= 0, \text{ elsewhere },$$

show that the characteristic function of $X + Y$ is equal to the product of the characteristic functions of X and Y. Comment on the result.

Hint. See remark to Theorem 6·22, page 6·81.

15. Let $K(t_1, t_2) = \log_e M(t_1, t_2)$ where $M(t_1, t_2)$ is the m.g.f. of X and Y. Show that :

$$\frac{\partial K(0,0)}{\partial t_1} = E(X)\;;\quad \frac{\partial^2 K(0,0)}{\partial t_1^{\,2}} = \text{Var}\,X\;;\quad \frac{\partial^2 K(0,0)}{\partial t_1\,\partial t_2} = \text{Cov}\,(X, Y)$$

[Delhi Univ. B.A.(Stat. Hons.), Spl. Course 1987]

OBJECTIVE TYPE QUESTIONS

I. Comment on the following, giving examples, if possible :

(i) M.g.f. of a r.v. always exist.

(ii) Characteristic function of a r.v. always exists.

(iii) M.g.f. is not affected by change of origin or/and scale.

(iv) $\phi_{X+Y}(t) = \phi_X(t)\,.\,\phi_Y(t)$ implies X and Y *are independent.*

(v) $\phi_X(t) = \phi_Y(t)$ implies X and Y have the same distribution.

(vi) $\phi(0) = 1$ and $|\phi(t)| \le 1$.

(vii) Variance of a r.v. is 5 and its mean does not exist.

(ix) It is possible to find a r.v. whose first k moments exist but $(k+1)^{th}$ moment does not exist.

(x) If a r.v. X has a symmetrical distribution about origin then

(a) $\phi_X(t)$ is even valued function of t.

(b) $\phi_X(t)$ is complex valued function of t.

II. *(a)* Can the following be the characteristic functions of any distribution ? Give reasons.

(*i*) $\log(1 + t)$, (*ii*) $\exp(-t^4)$, (*iii*) $1/(1 + t^4)$.

(b) Prove that $\phi(t) = \exp(-t^{\alpha})$, cannot be a characteristic function unless $\alpha = 2$.

III. State the relations, if any, between the following :

(i) $E(X^r)$ and $\phi_X(t)$.

(ii) $M_X(t)$ and $M_{X-a}(t)$, a being constant.

(iii) $M_X(t)$ and $M_{(X-a)/h}(t)$, a and h being constants.

(iv) $\phi_X(t)$ and p.d.f. of X.

(v) μ_r and μ_r'.

(vi) First four cumulants in terms of first four moments about mean.

IV. Let $X_1, X_2, ..., X_n$ be n *i.i.d.* (independent and identically distributed) r.v.'s with m.g.f. $M(t)$. Then prove that

$$M_{\bar{X}}(t) = [M(t/n)]^n,$$

where
$$\bar{X} = \sum_{i=1}^{n} X_i/n$$

V. If $X_1, X_2, ..., X_n$ are independent r.v.'s then prove that

$$M_{\sum\limits_{i=1}^{n} c_i X_i}(t) = \prod_{i=1}^{n} M_{X_i}(c_i t).$$

VI. Fill in the blanks :

(i) If $\int_{-\infty}^{\infty} |\phi_X(t)|\, dt < \infty$, then p.d.f. of X is given by ...

(ii) X_1 and X_2 are independent if and only if ... .

(Give result in terms of characteristic functions.)

(iii) If X_1 and X_2 are independent then

$$\phi_{X_1 - X_2}(t) = \ldots$$

(iv) $\phi(t)$ is ... defined and is ... for all t in $(-\infty, \infty)$.

VII. Examine critically the following statements :

(a) Two distributions having the same set of moments are identical.

(b) The characteristic function of a certain non-degenerate distribution is e^{-t^3}.

6·13. Chebychev's Inequality. The role of standard deviation as a parameter to characterise variance is precisely interpreted by means of the well known Chebychev's inequality. The theorem discovered in 1853 was later on discussed in 1856 by Bienayme.

Theorem 6·31. *If X is a random variable with mean μ and variance σ^2, then for any positive number k, we have*

$$P\left\{|X-\mu| \geq k\sigma\right\} \leq 1/k^2 \qquad \ldots(6·73)$$

or
$$P\left\{|X-\mu| < k\sigma\right\} \geq 1-(1/k^2) \qquad \ldots(6·73\,a)$$

Proof. Case (i). *X is a continuous r.v. By def.,*

$$\sigma^2 = \sigma_X^2 = E[X-E(X)]^2 = E[X-\mu]^2$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx, \text{ where } f(x) \text{ is p.d.f. of } X.$$

$$= \int_{-\infty}^{\mu - k\sigma} (x-\mu)^2 f(x) dx + \int_{\mu - k\sigma}^{\mu + k\sigma} (x-\mu)^2 f(x)\, dx + \int_{\mu + k\sigma}^{\infty} (x-\mu)^2 f(x)\, dx$$

$$\geq \int_{-\infty}^{\mu - k\sigma} (x-\mu)^2 f(x)\, dx + \int_{\mu + k\sigma}^{\infty} (x-\mu)^2 f(x)\, dx \qquad \ldots(*)$$

We know that :

$$x \leq \mu - k\sigma \text{ and } x \geq \mu + k\sigma \quad \Leftrightarrow \quad | x - \mu | \geq k\sigma \qquad \ldots(**)$$

Substituting in (*), we get

$$\therefore \qquad \sigma^2 \geq k^2\sigma^2 \left[\int_{-\infty}^{\mu - k\sigma} f(x)\, dx + \int_{\mu + k\sigma}^{\infty} f(x)\, dx \right]$$

$$= k^2 \sigma^2 \left[P\,(X \leq \mu - k\sigma) + P\,(X \geq \mu + k\sigma) \right] \qquad \text{[From (**)]}$$

$$= k^2 \sigma^2 .\ P\,(| X - \mu | \geq k\sigma) \qquad \text{[From (**)]}$$

$$\Rightarrow \quad P\,(| X - \mu | \geq k\sigma) \leq 1/k^2, \qquad \ldots(***)$$

which establishes (6·73)

Also since

$P\,\{| X - \mu | \geq k\,\sigma\} + P\,\{| X - \mu | < k\,\sigma\} = 1$, we get

$P\,\{| X - \mu | < k\,\sigma\} = 1 - P\,\{| X - \mu | \geq k\,\sigma\} \geq 1 - \{1/k^2\}$ [From (***)]

which establishes (6·73*a*).

Case (ii). In case of discrete random variable, the proof follows exactly similarly on replacing integration by summation.

Remark. In particular, if we take $k\,\sigma = c > 0$, then (*) and (**) give respectively

$$P\left\{| X - \mu | \geq c \right\} \leq \frac{\sigma^2}{c^2} \text{ and } P\left\{| X - \mu | < c\right\} \geq 1 - \frac{\sigma^2}{c^2}$$

$$\left.\begin{aligned} \Rightarrow \quad & P\left\{| X - E\,(X) | \geq c \right\} \leq \frac{\text{Var}\,(X)}{c^2} \\ \text{and} \quad & P\left\{| X - E\,(X) | < c\right\} \geq 1 - \frac{\text{Var}\,(X)}{c^2} \end{aligned}\right\} \qquad \ldots(6\cdot73\,b)$$

6·13·1. Generalised Form of Bienayme–Chebychev's Inequality. *Let $g\,(X)$ be a non-negative function of a random variable X. Then for every $k > 0$, we have*

$$P\,\{ g\,(X) \geq k \} \leq \frac{E\,\{ g\,(X) \}}{k} \qquad \ldots(6\cdot74)$$

[Bangalore Univ. B.Sc., 1992]

Proof. Here we shall prove the theorem for continuous random variable. The proof can be adapted to the case of discrete random variable on replacing integration by summation over the given range of the variable.

Let S be the set of all X where $g\,(X) \geq k$, *i.e.*,

$$S = \{x : g(x) \geq k\}$$

then $$\int_S dF(x) = P(X \in S) = P[g(X) \geq k], \quad \text{...(*)}$$

where $F(x)$ is the distribution function of X.

Now $$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, dF(x) \geq \int_S g(x)\, dF(x)$$

$$\geq k \, . \, P[g(X) \geq k] \quad [\because \text{ on } S, g(X) \geq k \text{ and using (*)}]$$

$$\Rightarrow \quad P[g(X) \geq k] \leq \frac{E[g(X)]}{k}$$

Remarks 1. If we take $g(X) = \{X - E(X)\}^2 = \{X - \mu\}^2$ and replace k by $k^2 \sigma^2$ in (6·74), we get

$$P\left\{(X-\mu)^2 \geq k^2 \sigma^2\right\} \leq \frac{E(X-\mu)^2}{k^2 \sigma^2} = \frac{\sigma^2}{k^2 \sigma^2} = \frac{1}{k^2}$$

$$\Rightarrow \quad P\left\{|X - \mu| \geq k\sigma\right\} \leq 1/k^2, \quad \text{(6·74 a)}$$

which is Chebychev's inequality.

2. *Markov's Inequality.* Taking $g(X) = |X|$ in (6·74) we get, for any $k > 0$

$$P[|X| \geq k] \leq \frac{E|X|}{k}, \quad \text{...(6·75)}$$

which is Markov's inequality.

Rather, taking $g(X) = |X|^r$ and replacing k by k^r in (6·74), we get a more generalised form of Markov's inequality, *viz.*,

$$P[|X|^r \geq k^r] \leq \frac{E|X|^r}{k^r} \quad \text{...(6·75a)}$$

3. If we assume the existence of only second-order moments of X, then we cannot do better than Chebychev's inequality (6·73). However, we can sometimes improve upon the results of Chebychev's inequality if we assume the existence of higher order moments. We give below (without proof) one such inequality which assumes the existence of moments of 4th order.

Theorem 6·31*a*. $E|X|^4 < \infty$, $E(X) = 0$ *and* $E(X^2) = \sigma^2$

$$P\{|X| > k\sigma\} \geq \frac{\mu_4 - \sigma^4}{\mu_4 + \sigma^4 k^4 - 2k^2 \sigma^4} \quad \text{...(6·76)}$$

If $X \sim U[0, 1]$, [*c.f.* Chapter 8], with p.d.f. $p(x) = 1, 0 < x < 1$ and $= 0$, otherwise, then

$$E(X^r) = 1/(r+1); \ (r = 1, 2, 3, 4)$$

$$E(X) = 1/2, \ E(X^2) = 1/3, \ E(X^3) = 1/4, \ E(X^4) = 1/5 \quad \text{...(*)}$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = 1/12$$

$$\mu_4 = E(X - \mu)^4 = E(X - \tfrac{1}{2})^4 = 1/80$$

[On using binomial expansion with (*)]

Chebychev's inequality (6·73a) with $k = 2$ gives :

$$P\left[\left|X - \tfrac{1}{2}\right| < 2\,\frac{1}{\sqrt{12}}\right] \geq 1 - \tfrac{1}{4} = 0{\cdot}75$$

With $k = 2$, (6·76) gives:

$$P\left[\left|X - \tfrac{1}{2}\right| > 2\frac{1}{\sqrt{12}}\right] \leq \frac{1/80 - 1/44}{1/80 + 1/9 - 1/8} = \frac{4}{49}$$

$$\Rightarrow \quad P\left[\left|X - \tfrac{1}{2}\right| \leq 2\frac{1}{\sqrt{12}}\right] \geq 1 - \frac{4}{49} = \frac{45}{49} = 0{\cdot}92,$$

which is a much better lower bound than the lower bound given by Chebychev's inequality.

6·14. Convergence in probability. We shall now introduce a new concept of convergenee, *viz.*, convergence in probability or stochastic convergence which is defined as follows:-

A sequence of random variables $X_1, X_2, ..., X_n, ...$ is said to converge in probability to a constant a, if for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - a| < \varepsilon) = 1 \qquad \text{...(6·77)}$$

or its equivalent

$$\lim_{n \to \infty} P(|X_n - a| \geq \varepsilon) = 0 \qquad \text{...(6·77a)}$$

and we write

$$X_n \xrightarrow{p} a \text{ as } n \to \infty \qquad \text{...(6·77b)}$$

If there exists a random variable X such that $X_n - X \xrightarrow{p} a$ as $n \to \infty$, then we say that the given sequence of random variables *converges in probability to the random variable X*.

Remark. 1. If a sequence of constants $a_n \to a$ as $n \to \infty$, then regarding the constant as a random variable having a one-point distribution at that point, we can say that as $a_n \xrightarrow{p} a$ as $n \to \infty$.

2. Although the concept of convergence in probability is basically different from that of ordinary convergence of sequence of numbers, it can be easily verified that the following simple rules hold for convergence in probability as well.

If $X_n \xrightarrow{p} \alpha$ and $Y_n \xrightarrow{p} \beta$ as $n \to \infty$, then

(*i*) $X_n \pm Y_n \xrightarrow{p} \alpha \pm \beta$ as $n \to \infty$

(*ii*) $X_n Y_n \xrightarrow{p} \alpha\beta$ as $n \to \infty$

(*iii*) $\dfrac{X_n}{Y_n} \xrightarrow{p} \dfrac{\alpha}{\beta}$ as $n \to \infty$, provided $\beta \neq 0$.

6·14·1. *(Chebychev's Theorem).* As an immediate consequence of Cheby-chev's inequality, we have the following theorem and convergence in probability.

"If $X_1, X_2, ..., X_n$ is a sequence of random variables and if mean μ_n and standard deviation σ_n of X_n exists for all n and if $\sigma_n \to -$ as $n \to \infty$, then

$$X_n - \mu_n \xrightarrow{p} 0 \text{ as } n \to \infty$$

Proof. We know, for any $\varepsilon > 0$

$$P\left\{ \left| X_n - \mu_n \right| \geq \varepsilon \right\} \leq \frac{\sigma_n^2}{\varepsilon^2} \to 0 \text{ as } n \to \infty$$

Hence $\quad X_n - \mu_n \xrightarrow{p} 0$ as $n \to \infty$

6·15. Weak Law of Large Numbers. *Let $X_1, X_2, ..., X_n$ be a sequence of random variables and $\mu_1, \mu_2, ..., \mu_n$ be their respective expectations and let*

$$B_n = \text{Var}(X_1 + X_2 + \ldots + X_n) < \infty$$

$$\text{Then } P\left\{ \left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{\mu_1 + \mu_2 + \ldots + \mu_n}{n} \right| < \varepsilon \right\} \geq 1 - \eta \qquad \ldots(6{\cdot}78)$$

for all $n > n_0$, where ε and η are arbitrary small positive numbers, provided

$$\lim_{n \to \infty} \frac{B_n}{n^2} \to 0$$

Proof. Using Chebychev's Inequality (6·73*b*), to the random variable $(X_1 + X_2 + \ldots + X_n)/n$, we get for any $\varepsilon > 0$,

$$P\left\{ \left| \left(\frac{X_1 + X_2 + \ldots + X_n}{n} \right) - \left(E\,\frac{X_1 + X_2 + \ldots + X_n}{n} \right) \right| < \varepsilon \right\} \geq 1 - \frac{B_n}{n^2 \varepsilon^2},$$

$$\left[\text{since Var}\left(\frac{X_1, X_2 + \ldots + X_n}{n} \right) = \frac{1}{n^2} \text{Var}(X_1 + X_2 + \ldots + X_n) = \frac{B_n}{n^2} \right]$$

$$\Rightarrow \quad P\left\{ \left| \frac{X_1, X_2 + \ldots + X_n}{n} - \frac{\mu_1 + \mu_2 + \ldots + \mu_n}{n} \right| < \varepsilon \right\} \geq 1 - \frac{B_n}{n^2 \varepsilon^2}.$$

So far, nothing is assumed about the behaviour of B_n for indefinitely increasing values of n. Since ε is arbitrary, we assume $\dfrac{B_n}{n^2 \varepsilon^2} \to 0$, as n becomes indefinitely large. Thus, having chosen two arbitary small positive numbers ε and η, number n_0 can be found so that the inequality

$$\frac{B_n}{n^2 \varepsilon^2} < \eta,$$

will hold for $n > n_0$. Consequently, we shall have

$$P\left\{ \left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{\mu_1 + \mu_2 + \ldots + \mu_n}{n} \right| < \varepsilon \right\} \geq 1 - \eta$$

for all $n > n_0(\varepsilon, \eta)$.

This conclusion leads to the following important result, known as the (*Weak*) Law of Large Numbers:

"With the probability approaching unity or certainty as near as we please, we may expect that the arithmetic mean of values actually assumed by n random variables will differ from the arithmetic mean of their expectations by less than any given number, however small, provided the number of variables can be taken sufficently large and provided the condition

$$\frac{B_n}{n^2} \to 0 \text{ as } n \to \infty$$

is fulfilled".

Remarks. 1. Weak law of large numbers can also be stated as follows:

$$\bar{X}_n \xrightarrow{p} \bar{\mu}_n$$

provided $\frac{B_n}{n^2} \to 0$ as $n \to \infty$, symbols having their usual meanings.

2. For the existence of the law we assume the following conditions:

(*i*) $E(X_i)$ exists for all i, (*ii*) $B_n = \text{Var}(X_1 + X_2 + \ldots + X_n)$ exists, and

(*iii*) $B_n/n^2 \to 0$ as $n \to \infty$.

Condition (*i*) is necessary, without it the law itself cannot be stated. But the conditions (*ii*) and (*iii*) are not necessary, (*iii*) is however a sufficient condition.

3. If the variables $X_1, X_n, \ldots, X_n$ are independent and identically distributed, *i.e.*, if $E(X_i) = \mu$ (say), and $\text{Var}(X_i) = \sigma^2$ (say) for all $i = 1, 2, \ldots, n$ then

$$B_n = \text{Var}(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i)$$

the convariance terms vanish, since variables are independent.

$$\therefore \qquad B_n = n\sigma^2 \qquad \ldots(*)$$

Hence $$\lim_{n \to \infty} \frac{B_n}{n^2} = \lim_{n \to \infty} (\sigma^2/n) = 0$$

Thus, the law of large number holds for the sequence $\{X_n\}$ of *i.i.d. r.v.'s* and we get

$$P\left\{ \left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \mu \right| < \varepsilon \right\} > 1 - \eta \ \forall \ n > n_0$$

i.e., $P\{|\bar{X}_n - \mu| < \varepsilon\} \to 1$ as $n \to \infty$

$\Rightarrow$ $P\{|\bar{X}_n - \mu| \geq \varepsilon\} \to 0$ as $n \to \infty$

where $\bar{X}_n$ is the mean of the n random variables $X_1, X_2, \ldots, X_n$. This result implies that $\bar{X}_n$ converges in *probability* to μ, *i.e.*,

$$\bar{X}_n \xrightarrow{p} \mu$$

Note. If $\overline{X}_n$ is the mean of n *i.i.d.* r.v.'s $X_1, X_2, \ldots, X_n$ with

$E(X_i) = \mu$; $\text{Var}(X_i) = \sigma^2$, then [On using (*)]

$$E(\overline{X}_n) = \mu \text{ ; and } \text{Var}(\overline{X}_n) = \text{Var}\left(\sum_{i=1}^{n} x_i / n\right) \qquad \ldots(6\cdot80)$$

Theorem 6·32. *If the variables are uniformly bounded then the condition,*

$$\lim_{n \to \infty} \frac{B_n}{n^2} = 0$$

is necessary as well as sufficient for WLLN to hold.

Proof. Let $\xi_i = X_i - a_i$, where $E(X_i) = a_i$; then $E(\xi_i) = 0$, $(i = 1, 2, \ldots, n)$.

Since X_i's are uniformly bounded, there exists a positive number $c < \infty$ such that $|\xi_i| < c$.

If $\quad p = P[\,|\xi_1 + \xi_2 + \ldots + \xi_n| \le n\varepsilon\,]$

then $\quad 1 - p = P[\,|\xi_1 + \xi_2 + \ldots + \xi_n| > n\varepsilon\,]$

Let $\quad U_n = \xi_1 + \xi_2 + \ldots + \xi_n$,

then $\quad E(U_n) = \sum_{i=1}^{n} E(\xi_i) = 0$

and $\quad \text{Var}(U_n) = E(U_n^2) = B_n$ (say).

$$\Rightarrow \quad B_n = \int_0^\infty U_n^2\, dF(U_n), \text{ where } F(U_n) \text{ is d.f. of } U_n.$$

$$= \int_{U_n^2 \le n^2\varepsilon^2} U_n^2\, dF + \int_{U_n^2 > n^2\varepsilon^2} U_n^2\, dF$$

$$\le n^2\varepsilon^2 \int_{|U_n| \le n\varepsilon} dF + n^2 c^2 \int_{|U_n| > n\varepsilon} dF$$

$$\le n^2\varepsilon^2 p + n^2 c^2 (1 - p)$$

$$\therefore \quad \frac{B_n}{n^2} \le \varepsilon^2 p + c^2 (1 - p)$$

If the law of large numbers holds,

$$1 - p = P[\,|\xi_1 + \xi_2 + \ldots + \xi_n| > n\varepsilon\,] \to 0 \text{ as } n \to \infty.$$

Hence as $n \to \infty$, $(1 - p) \to 0$, and

$\dfrac{B_n}{n^2} < \varepsilon^2 p + c^2 \delta$, ε and δ being arbitrarily small positive numbers.

Hence $\quad \dfrac{B_n}{n^2} \to 0$ as $n \to \infty$.

6·15·1. Bernoulli's Law of Large Numbers. Let there be n trials of an event, each trial resulting in a success or failure. If X is the number of successes in n trials with constant probability p of success for each trial, then $E(X) = np$ and $\text{Var}(X) = npq$, $q = 1 - p$. The variable X/n represents the proportion of successes or the relative frequency of successes, and

$$E(X/n) = \frac{1}{n} E(X) = p, \text{ and } \text{Var}(X/n) = \frac{1}{n^2} \text{Var}(X) = \frac{pq}{n}$$

Then

$$P\left\{ \left| \frac{X}{n} - p \right| < \varepsilon \right\} \to 1 \text{ as } n \to \infty \quad \ldots(6·81)$$

$$\Rightarrow \quad P\left\{ \left| \frac{X}{n} - p \right| \geq \varepsilon \right\} \to 0 \text{ as } n \to \infty \quad \ldots(6·81\ a)$$

tor any assigned $\varepsilon > 0$. This implied that (X/n) converges in probability to p as $n \to \infty$.

Proof. Applying Chebychev's Inequality [Form (6·73 b)] to the variable X/n, we get for any $\varepsilon > 0$;

$$P\left\{ \left| \frac{X}{n} - E\left(\frac{X}{n}\right) \right| \geq \varepsilon \right\} \frac{\text{Var}(X/n)}{\varepsilon^2}$$

$$\Rightarrow \quad P\left\{ \left| \frac{X}{n} - p \right| < \varepsilon \right\} \to 1 \text{ as } n \to \infty$$

since the maximum value of pq is at $p = q = 1/2$ *i.e.*, max $(p\, q) = 1/4$ *i.e.*, $pq \leq 1/4$.

Since ε is arbitrary, we get

$$P\left\{ \left| \frac{X}{n} - p \right| \geq \varepsilon \right\} \to 0 \text{ as } n \to \infty$$

$$\Rightarrow \quad P\left\{ \left| \frac{X}{n} - p \right| < \varepsilon \right\} \to 1 \text{ as } n \to \infty$$

6·15·2. MorkofT's Theorem. *The law of large numbers holds if for some* $\delta > 0$, *all the mathematical expectations*

$$E(|X_i|^{1+\delta})\ ; i = 1, 2, \ldots \quad \ldots(6·82)$$

exist and are bounded.

6·15·3. Khinchin's Theorem. *If X_i's are identically and independently distributed random variables, the only condition necessary for the law of large numbers to hold is that $E(X_i)$; $i = 1, 2, \ldots$ should exist.*

Theorem 6·33. *Let $\{X_n\}$ be any sequence of r.v.'s. Write :*

$$Y_n = [S_n - E(S_n)]/n \text{ where } S_n = X_1 + X_2 + \ldots + X_n.$$

A necessary and sufficient condition for the sequence $\{X_n\}$ to satisfy the W.L.L.N. is that

$$E\left\{ \frac{Y_n^2}{1 + Y_n^2} \right\} \to 0 \text{ as } n \to \infty. \quad \ldots(6·83)$$

Proof. *If Part:* Let us assume that (6·83) holds. We shall prove $\{X_n\}$ satisfies W.L.L.N.

For real numbers a, b ; $a \geq b > 0$ we have:

$$a \geq b \quad \Rightarrow \quad a + ab \geq b + ab \qquad \text{...(*)}$$

Let us define the event $A = \{ \, |Y_n| \geq \varepsilon \, \}$.

$$w \in A_X \quad \Rightarrow \quad \left| Y_n \right| \geq \varepsilon \quad \Rightarrow \quad \left| Y_n \right|^2 \geq \varepsilon^2 > 0$$

$\therefore$ Taking $a = Y_n^2$ and $b = \varepsilon^2$ in (*), we define another event B as follows:

$$B_n = \left\{ \left(\frac{Y^{n^2}}{1 + Y_n^2} \right) \left(\frac{1 + \varepsilon^2}{\varepsilon^2} \right) \geq 1 \right\} = \left\{ \frac{Y_n^2}{1 + Y_n^2} \geq \frac{\varepsilon^2}{1 + \varepsilon^2} \right\}$$

Since $w \in A \quad \Rightarrow \quad w \in B, \quad A \subseteq B \quad \Rightarrow \quad P(A) \leq P(B)$

$$\therefore \qquad P\left[\, \left| Y_n \right| \geq \varepsilon \, \right] \leq P\left[\frac{y_n^2}{1 + Y_n^2} \frac{\varepsilon^2}{1 + \varepsilon^2} \right]$$

$$\leq \frac{E\left\{ Y_n/(1 + Y_n^2) \right\}}{\varepsilon^2/(1 + \varepsilon^2)}$$

[By Markov's Inequality (6·75)]

$$\to 0 \text{ as } n \to \infty \qquad \text{[By assumption (6·83)]}$$

$$\therefore \qquad P\left[\, \left| Y_n \right| \geq \varepsilon \, \right] \to 0 \text{ as } n \to \infty$$

$$\Rightarrow \qquad \lim_{n \to \infty} P\left[\left| \frac{S_n - E(S_n)}{n} \right| \geq \varepsilon \right] \to 0$$

$\Rightarrow$ WLLN holds for the sequence $\{X_n\}$ of r.v.'s.

Conversely, if $\{X_n\}$ satisfies WLLN, we shall establish (6·83). Let us assume that X_i's are continuous and let Y_n have p.d.f. $f_n(y)$. Then

$$E\left\{ \frac{Y_n^2}{1 + Y_n^2} \right\} = \int_{-\infty}^{\infty} \frac{y^2}{1 + y^2} \cdot f_n(y) \, dy$$

$$= \left(\int_A + \int_{A^c} \right) \frac{y^2}{1 + y^2} f_n(y) \, dy$$

where $\qquad A = \{ \, |Y| \geq \varepsilon \, \}$ and $A^c = \{ \, |y| < \varepsilon \, \}$

$$E\left(\frac{Y_n^2}{1+Y_n^2}\right) \le \int_A 1.f_n(y)\,dy + \int_{A^c} y^2.f_n(y)\,dy \;\Big|\because \frac{y^2}{1+y^2} < 1 \text{ and } \frac{y^2}{1+y^2} < y^2\Big|$$

$$\le P(A) + \varepsilon^2 \int_{A^c} f_n(y)\,dy \qquad (\because \text{On } A^c : |y| < \varepsilon)$$

$$= P(A) + \varepsilon^2 . P(A^c)$$

$$\le P(A) + \varepsilon^2 \qquad (\because P(A^c) < 1)$$

$$\Rightarrow E\left[\frac{Y_n^2}{1+Y_n^2}\right] \le P\{|Y_n| \ge \varepsilon\} + \varepsilon^2 \qquad \ldots(**)$$

But since $\{X_n\}$ satisfies WLLN, we have :

$$\lim_{n\to\infty} P[|Y_n| \ge \varepsilon] \to 0$$

and since ε^2 is arbitrarily small positive number, we get on taking limits in (**).

$$\lim_{n\to\infty} E\left[\frac{Y_n^2}{1+Y_n^2}\right] \to 0$$

Corollary. Let $X_1, X_2, \ldots, X_n$ be sequence of independent r.v.'s such that Var $(X_i) < \infty$ for $i = 1, 2, \ldots$ and

$$\frac{B_n}{n^2} = \frac{\text{Var}\left(\sum\limits_{i=1}^{n} X_i\right)}{n^2} = \frac{\text{Var}(S_n)}{n^2} \to 0 \text{ as } n \to \infty$$

Then WLLN holds.

Proof. We have :

$$\frac{Y_n^2}{1+Y_n^2} \le Y_n^2 = \left[\frac{S_n - E(S_n)}{n}\right]^2$$

$$\Rightarrow \qquad E\left[\frac{Y_n^2}{1+Y_n^2}\right] \le \frac{1}{n^2}\cdot E[S_n - E(S_n)]^2 = \frac{\text{Var } S_n}{n^2} = \frac{B_n}{n^2}$$

$$\therefore \qquad \lim_{n\to\infty} E\left[\frac{Y_n^2}{1+Y_n^2}\right] \le \lim_{n\to\infty} \frac{B_n}{n^2} \to 0 \qquad \text{(By assumption)}$$

Hence by the above theorem WLLN holds for the sequence $\{X_n\}$ of r.v.'s.

Remark. The result of Theorem 6·33 holds even if $E(X_i)$ does not exist. In this case we simply define $Y_n = [S_n/n]$ rather than $[S_n - E(S_n)]/n$.

Example 6·48. *A symmetric die is thrown 600 times. Find the lower bound for the probability of getting 80 to 120 sixes.*

Solution. Let S be total number of successes.

Then

$$E(S) = np = 600 \times \frac{1}{6} = 100$$

$$V(S) = npq = 600 \times \frac{1}{6} \times \frac{5}{6} = \frac{500}{6}$$

Using Chebychev's inequality, we get

$$P[\,|S - E(S)| < k\sigma\,] \geq 1 - \frac{1}{k^2}$$

$$\Rightarrow \qquad P\{\,|S - 100| < k\sqrt{500/6}\,\} \geq 1 - \frac{1}{k^2}$$

$$\Rightarrow \qquad P\{\,100 - k\sqrt{500/6} < S < 100 + k\sqrt{500/6}\,\} \geq 1 - \frac{1}{k^2}$$

Taking $\quad k = \dfrac{20}{\sqrt{500/6}}$, we get

$$P(80 \leq S \leq 120) \geq 1 - \frac{1}{400 \times (6/500)} = \frac{19}{24}$$

Example 6·49. *Use Chebychev's inequality to determine how many times a fair coin must be tossed in order that the probability will be at least 0·90 that the ratio of the observed number of heads to the number of tosses will lie between 0·4 and 0·6.*

[Madras Univ. B.Sc (Stat.)Oct 1991; Delhi Univ. B.Sc. (Stat Hons.) 1989]

Solution. As in the proof of Bernoulli's Law of Large Numbers, we get for any $\varepsilon > 0$,

$$P\left\{\left|\frac{X}{n} - p\right| \geq \varepsilon\right\} \leq \frac{1}{4n\varepsilon^2}$$

$$\Rightarrow \qquad P\left\{\left|\frac{X}{n} - p\right| < \varepsilon\right\} \geq 1 - \frac{1}{4n\varepsilon^2}$$

Since $p = 0{\cdot}5$ (as the coin is unbaised) and we want the proportion of successes X/n to lie between 0·4 and 0·6, we have

$$\left|\frac{X}{n} - p\right| \leq 0{\cdot}1$$

Thus choosing $\varepsilon = 0{\cdot}1$, we have

$$P\left\{\left|\frac{X}{n} - p\right| < 0{\cdot}1\right\} \geq 1 - \frac{1}{4n(0{\cdot}1)^2} = 1 - \frac{1}{0{\cdot}04\,n}$$

Since we want this probability to be 0·9, we fix

$$1 - \frac{1}{0{\cdot}04\,n} = 0{\cdot}90$$

$$\Rightarrow \qquad 0{\cdot}10 = \frac{1}{0{\cdot}04n}$$

$$\Rightarrow \qquad n = \frac{1}{0{\cdot}10 \times 0{\cdot}04} = 250$$

Hence the required number of tosses is 250.

Example 6·50. *For geometric distribution $p(x) = 2^{-x}$; $x = 1, 2, 3, \ldots$ prove that Chebychev's inequality gives*

$$P[\,|X - 2| \le \,] > \tfrac{1}{2}$$

while the actual probability is $\frac{15}{16}$. **[Nagpur Univ. B.Sc. (Stat.), 1989]**

Solution. $E(X) = \sum\limits_{x=1}^{\infty} \dfrac{x}{2^x} = \dfrac{1}{2} + \dfrac{2}{2^2} + \dfrac{3}{2^3} + \dfrac{4}{2^4} + \ldots$

$$= \tfrac{1}{2}(1 + 2A + 3A^2 + 4A^3 + \ldots), \quad (A = 1/2)$$

$$= \tfrac{1}{2}(1 - A)^{-2} = 2$$

$$E(X^2) = \sum_{x=1}^{\infty} \frac{x^2}{2^x} = \frac{1}{2^2} + \frac{4}{2^3} + \frac{9}{2^4} + \ldots$$

$$= \tfrac{1}{4}[1 + 4A + 9A^2 + \ldots] = \tfrac{1}{4}(1 + A)(1 - A)^{-3} = 6$$

[See Example 6·17]

$$\therefore \qquad \text{Var}(X) = E(X^2) - [E(X)]^2 = 6 - 4 = 2$$

Using Chebychev's inequality, we get

$$P\{|X - E(X)| \le k\sigma\} > 1 - \frac{1}{k^2}$$

With $k = \sqrt{2}$, we get

$$P\{|X - 2| \le \sqrt{2}\,.\sqrt{2}\} > 1 - \tfrac{1}{2} = \tfrac{1}{2}$$

$$\Rightarrow \qquad P\{|X - 2| \le 2\} > \tfrac{1}{2}$$

And the actual probability is given by

$$P\{|X - 2| \le 2\} = P\{0 \le X \le 4\} = P\{X = 1, 2, 3 \text{ or } 4\}$$

$$= \tfrac{1}{2} + (\tfrac{1}{2})^2 + (\tfrac{1}{2})^3 + (\tfrac{1}{2})^4 = \frac{15}{16}$$

Example 6·51. *Does there exist a variate X for which*

$$P[\mu_x - 2\sigma \le X \le \mu_x + 2\sigma] = 0{\cdot}6 \qquad \ldots(*)$$

[Delhi Univ. B.Sc.(Maths Hons.) 1983]

Solution. We have :

$$P[\mu_x - 2\sigma \le X \le \mu_x + 2\sigma] = P[|X - \mu_x| \le 2\sigma]$$

$$\ge 1 - \frac{1}{4} = 0{\cdot}75 \qquad \text{(Using Chebychev's Inequality)}$$

Since lower bound for the probability is 0·75, there does not exist a r.v. X for which (*) holds.

Example 6·52. *(a) For the discrete variate with density*

$$f(x) = \frac{1}{8}\, I_{(-1)}(x) + \frac{6}{8}\, I_{(0)}(x) + \frac{1}{8}\, I_{(1)}(x)$$

evaluate $P[|X - \mu_x| \geq 2\,\sigma_x]$. **[Delhi Univ. B.Sc.(Maths Hons.) 1989]**

(b) Compare this result with that obtained on using Chebychev's inequality.

Hint. *(a)* Here X has the probability distribution :

x :	-1	0	1
$p(x)$:	$1/8$	$6/8$	$1/8$

$$\therefore\ E(X) = -1 \times \frac{1}{8} + 1 \times \frac{1}{8} = 0$$

$$EX^2 = 1 \times \frac{1}{8} + 1 \times \frac{1}{8} = \frac{1}{4}$$

$\therefore\quad \text{Var}(X) = E(X^2) - [E(X)]^2 = 1/4 \quad \Rightarrow \quad \sigma_x = 1/2$

$$P[|X - \mu_x| \geq 2\,\sigma_x] = P[|X| \geq 1] = 1 - P(|X| < 1)$$
$$= 1 - P[-1 < X < 1] = 1 - P(X = 0) = 1/4$$

(b) $\qquad P[|X - \mu_x| \geq 2\,\sigma_x] \leq \frac{1}{4}$ (By Chebychev's Inequality)

In this case both results are same.

Note. This example shows that, in general, Chebychev's inequality cannot be improved.

Example 6·53. *Two unbiased dice are thrown. If X is the sum of the numbers showing up, prove that*

$$P(|X - 7| \geq 3) \leq \frac{35}{54}.$$

Compare this with the actual probability. **(Karnataka Univ. B.Sc., 1988)**

Solution. The probability distribution of the r.v. X (the sum of the numbrs on the two dice) is as given below :

X	Favourable cases (distinct)	Probability (p)
2	(1, 1)	1/36
3	(1, 2), (2, 1)	2/36
4	(1, 3), (3, 1), (2, 2)	3/36
5	(1, 4), (4, 1), (2, 3), (3, 2)	4/36
6	(1, 5), (5, 1), (2, 4), (4, 2), (3, 3)	5/36
7	(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)	6/36
8	(2, 6), (6, 2), (3, 5), (5, 3), (4, 4)	5/36
9	(3, 6), (6, 3), (4, 5), (5, 4)	4/36
10	(4, 6), (6, 4), (5, 5)	3/36
11	(5, 6), (6, 5)	2/36
12	(6, 6)	1/36

$$E(X) = \sum_x p.x$$
$$= \frac{1}{36}(2 + 6 + 12 + 20 + 30 + 42 + 40 + 36 + 30 + 22 + 12)$$
$$= \frac{1}{36}(252) = 7$$
$$E(X^2) = \sum_x p.x^2$$
$$= \frac{1}{36}[4 + 18 + 48 + 100 + 180 + 294 + 320 + 324 + 300 + 242 + 144]$$
$$= \frac{1}{36}(1974) = \frac{1974}{36} = \frac{329}{6}$$
$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{329}{6} - (7)^2 = \frac{35}{6}$$

By Chebychev's inequality, for $k > 0$, we have

$$P(|X - \mu| \ge k) \le \frac{\text{Var}\,X}{k^2}$$
$$\Rightarrow \quad P(|X - 7| \ge 3) \le \frac{35/6}{9} = \frac{35}{54} \qquad \text{(Taking } k = 3)$$

Actual Probability :

$$P(|X - 7| \ge 3) = 1 - P(|X - 7| < 3$$
$$= 1 - P(4 < X < 10)$$
$$= 1 - [P(X = 5) + P(X = 6) + P(X = 7) \; P(X = 8) + P(X = 9)]$$
$$= 1 - \frac{1}{36}[4 \quad 5 + 6 + 5 + 4] = 1 - \frac{24}{36} = \frac{1}{3}$$

Example 6·54. *If X is the number scored in a throw of a fair die, show that the Chebychev's inequality gives*

$$P[|X - \mu| > 2\cdot5] < 0\cdot47,$$

where μ is the mean of X, while the actual probability is zero.

[Kerala Univ. B.Sc., Oct. 1989]

Solution. Here X is a random variable which takes the values 1,2,...,6, each with probability 1/6. Hence

$$E(X) = \tfrac{1}{6}(1 + 2 + \ldots + 6) = \tfrac{1}{6}\cdot\frac{6 \times 7}{2} = \frac{7}{2}$$
$$E(X^2) = \tfrac{1}{6}(1^2 + 2^2 + \ldots + 6^2) = \tfrac{1}{6}\,\frac{6 \times 7 \times 13}{6} = \frac{91}{6}$$
$$\therefore \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} = 2\cdot9167$$

For $k > 0$, Chebychev's inequality gives

$$P[\,|X-E(X)|>k\,]<\frac{\text{Var}\,X}{k^2}$$

Choosing $k = 2{\cdot}5$, we get

$$P[\,|X-\mu|>2{\cdot}5\,]<\frac{2{\cdot}9167}{6{\cdot}25}=0{\cdot}47$$

The actual probability is given by

$$p=P[\,|X-3{\cdot}5|>2{\cdot}5\,]$$

$$=P[\,X \text{ lies outside the limits } (3{\cdot}5-2{\cdot}5,\ 3{\cdot}5+2{\cdot}5),\ i.e.,\ (1,6)]$$

But since X is the numberon the dice when thrown, it cannot lie outside the limits of 1 and 6.

$$\therefore \qquad p=p(\varphi)=0$$

Example 6·55. *If the variable X_p assumes the value $2^{p-2\log p}$ with probability 2^{-p} ; $p = 1, 2, \ldots$, examine if the law of large numbers holds in this case.*

Solution. Putting $p = 1, 2, 3, \ldots$, we get

$$2^{1-2\log 1},\ 2^{2-2\log 2},\ 2^{3-2\log 3}, \ldots$$

as the values of the variables $X_1, X_2, X_3 \ldots$respectively, and

$$\frac{1}{2}, \frac{1}{2^2}, \frac{1}{2^3}, \ldots$$

their corresponding probabilities. Therefore,

$$E(X_k)=2^{1-2\log 1}\cdot\frac{1}{2}+2^{2-2\log 2}\cdot\frac{1}{2^2}+\ldots=\sum_{p=1}^{\infty}2^{p-2\log p}\cdot 2^{-p}$$

$$=\sum_{p=1}^{\infty}\frac{1}{2^{2\log p}}$$

Let $\qquad U = 2^{2\log p}$, then

$$\log U = 2\log p\log 2=\log p\,.\,\log 2^2=\log 4\,.\,\log p=\log p^{\log 4}$$

$$\therefore \quad E(X_k)=\sum_{p=1}^{\infty}\frac{1}{p^{\log 4}}=1+\frac{1}{2^{\log 4}}+\frac{1}{3^{\log 4}}+\ldots$$

which is a convergent series.

$$\left[\because \sum_{n=1}^{\infty}\frac{1}{n^p} \text{ is convergent if and only if } p>1\right]$$

Therefore, the mathematical expectation of the variables $X_1, X_2, \ldots$, exists. Thus by Khinchin's theorem, the law of large numbers holds in this case.

Exmaple 6·56. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. variables with mean μ and variance σ^2 and as $n \to \infty$,*

$$(X_1^2+X_2^2+\ldots+X_n^2)/n \xrightarrow{p} c,$$

for some constant c; ($0 \le c \le \infty$). Find c. **[Delhi Univ. B.Sc. (Stat. Hons.), 1989]**

Solution. We are given $E(X_i) = \mu$; $\text{Var}(X_i) = \sigma^2$; $i = 1, 2, ..., n$.

$\therefore$ $E(X_i^2) = \text{Var}(X_i) + [E(X_i)]^2 = \sigma^2 + \mu^2$ (finite) ; $i = 1, 2, ..., n$.

Since $E(X_i^2)$ is finite; by Khinchines's Theorem WLLN holds for the sequence $\{X_i^2\}$ of i.i.d. r.v.'s so that

$$(X_1^2 + X_2^2 + \dots + X_n^2)/n \xrightarrow{p} E(X_i^2), \text{ as } n \to \infty$$

$$\Rightarrow \quad (X_1^2 + X_2^2 + \dots + X_n^2)/n \xrightarrow{p} \sigma^2 + \mu^2 = c, \text{ as } n \to \infty$$

Hence $\quad c = \sigma^2 + \mu^2$.

Example 6·57. *How large a sample must be taken in order that the probability will be at least 0·95 that $\overline{X}_n$ will be lie within 0·5 of μ . μ is unknown and* $\sigma = 1$. **[Delhi Univ. B.Sc. (Maths Hons.) 1988]**

Solution. We have: $E(\overline{X}_n) = \mu$ and $\text{Var}(\overline{X}_n) = \sigma^2/n$

[c.f. § 6·15, = n (6·80)]

Applying Chebychev's inequality to the r.v. $\overline{X}_n$ we get, for any $c > 0$

$$P[|\overline{X}_n - E(\overline{X}_n)| < c] \geq 1 - \frac{\text{Var}(\overline{X}_n)}{c^2}$$

$$\Rightarrow \quad P[|\overline{X}_n - \mu| < c] \geq 1 - \frac{\sigma^2}{n c^2} \quad ...(*)$$

We want n so that

$$P[|\overline{X}_n - \mu| < 0·5] \geq 0·95 \quad ...(**)$$

Comparing (*) and (**) we get :

$$c = 0·5 = 1/2 \quad \text{and} \quad 1 - \frac{\sigma^2}{n c^2} = 0·95 \quad \text{and} \quad \sigma = 1 \quad \text{(Given)}$$

$$\therefore \quad 1 - \frac{4}{n} = 0·95 \quad \Rightarrow \quad \frac{4}{n} = 0·05 = \frac{1}{20} \quad \Rightarrow \quad n = 80.$$

Hence $n \geq 80$.

Example 6·58. *(a) Let X_i assume that values i and –i with equal probabilities. Show that the law of large numbers cannot be applied to the independent variables $X_1, X_2, ...,$ i.e., X_i's.*

(b) If X_i can have only two values with equal probabilities i^{α} and $-i^{\alpha}$, show that the law of large numbers can be applied to the independent variables $X_1, X_2, ...,$ if $\alpha < \frac{1}{2}$.

Solution. (*a*) We have

$$P(X_i = i) = \tfrac{1}{2}, \quad P(X_i = -i) = \tfrac{1}{2}$$

$$E(X_i) = \tfrac{1}{2}(i) + \tfrac{1}{2}(-i) = 0 \; ; \; i = 1, 2, 3, \dots$$

$$V(X_i) = E(X_i^2) = \frac{i^2}{2} + \frac{i^2}{2} = i^2 \quad | \because E(X_i) = 0 \,| \quad ...(*)$$

$$B_n = V(X_1 + X_2 + \ldots + X_n) = V(X_1) + V(X_2) + \ldots + V(X_n)$$

$$= (1 + 2^2 + \ldots + n^2) = \frac{n(n+1)(2n+1)}{6} \qquad \text{...[From (*)]}$$

$\therefore \quad \frac{B_n}{n^2} \to \infty$ as $n \to \infty$. Hence we cannot draw any conclusion whether WLLN holds or not. Here, we need to apply further tests. (See Theorem 6·33 page 6·104)

(b) $E(X_i) = \frac{i^\alpha}{2} + \left(\frac{-i^\alpha}{2}\right) = 0$

$$E(X_i^2) = \frac{(i^\alpha)^2}{2} + \frac{(-i^\alpha)^2}{2} = i^{\alpha 2}$$

$$V(X_i) = E(X_i^2) - [E(X_i)]^2 = i^{2\alpha}$$

$$B_n = V(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} i^{2\alpha}$$

$$= 1^{2\alpha} + 2^{2\alpha} + \ldots + 2^{2\alpha} = \int_0^n x^{2a}\, dx$$

[From Euler Maclaurin's Formula]

$$= \left| \frac{x^{2\alpha+1}}{2\alpha+1} \right|_0^n = \frac{n^{2\alpha+1}}{2\alpha+1}$$

$$\therefore \quad \frac{B_n}{n^2} = \frac{n^{2\alpha-1}}{2\alpha+1} \to 0 \text{ if } 2\alpha - 1 < 0 \Rightarrow \alpha < \tfrac{1}{2}$$

Hence the result.

Example 6·59. Let $\{X_k\}$ be mutually independent and identically distributed random variables with mean μ and finite variance. If $S_n = X_1 + X_2 + \ldots + X_n$, prove that the law of large numbers does not hold for the sequence $\{S_n\}$.

Solution. The variables now are $S_1, S_2, \ldots, S_n$.

$$B_n = V(S_1 + X_2 + \ldots + S_n)$$

$$= V\{X_1 + (X_1 + X_2) + (X_1 + X_2 + X_3) + \ldots + (X_1 + X_2 + \ldots + X_n)\}$$

$$= V\{nX_1 + (n-1)X_2 + \ldots + 2X_{n-1} + X_n\}$$

$$= n^2 V(X_1) + (n-1)^2 V(X_2) + \ldots + 2^2 V(X_{n-1}) + 1^2 V(X_n)$$

(Covariance terms vanish since variables are independent.)

Let $V(X_i) = \sigma^2$ for all i.

(Since the variables are identically distributed.)

$$\therefore \quad B_n = (1^2 + 2^2 + \ldots + n^2)\sigma^2 + \frac{n(n+1)(2n+1)}{6}\sigma^2$$

$$\therefore \quad \frac{B_n}{n^2} = \frac{(n+1)(2n+1)}{6n} \cdot \sigma^2 \to \infty \text{ as } n \to \infty$$

Example 6·60. *Examine whether the week law of large numbers holds for the sequence* $\{X_k\}$ *of independent random variables defined as follows:*

$$P[X_k = \pm 2^k] = 2^{-(2k+1)}$$

$$P[X_k = 0] = 1 - 2^{-2k}$$ **[Delhi Univ. B.Sc. (Maths Hons.), 1988]**

Solution. We have

$$E(X_k) = 2^k\, 2^{-(2k+1)} + (-2^k) \,.\, 2^{-(2k+1)} + 0 \times (1 - 2^{-2k})$$

$$= 2^{-(2k+1)}\ (2^k - 2^k) = 0$$

$$E(X_k^2) = (2^k)^2 \,.\, 2^{-(2k+1)} + (-2^k)^2 \,.\, 2^{-(2k+1)} + 0^2 \times (1 - 2^{-2k})$$

$$= 2^{2k} \,.\, 2^{-(2k+1)} + 2^{k^2} \,.\, 2^{-(2k+1)}$$

$$= \tfrac{1}{2} + \tfrac{1}{2} = 1$$

$$\therefore \quad \text{Var}(X_k) = E(X_k^2) - E(X_k))^2 = 1 - 0 = 1$$

$$B_n = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}\ (X_i) \ :$$

[$\because$ X_i's $(i = 1, 2, ..., n)$ are independent]

$$= \sum_{i=1}^{n} (1) = n$$

$$\therefore \quad \lim_{n\to\infty} \frac{B_n}{n^2} = \lim_{n\to\infty} \frac{1}{n} \to 0$$

Hence (Weak) Law of large numbers, holds for the sequence of independent r.v.'s $\{X_k\}$.

Example 6·61. *Let* $X_1, X_2, \dots X_n$ *be jointly normal with* $E(X_i) = 0$ *and* $E(X_i^2) = 1$ *for all i and* $\text{Cov}(X_i, X_j) = \rho$ if $|j - i| = 1$ *and* $= 0$, *otherwise. Examine if WLLN holds for the sequence* $\{X_n\}$. ...(*)

Solution. We have:

$$\text{Var}(X_i) = (X_i^2) - [E(X_i)]^2 = 1, \quad (i = 1, 2, ..., n).$$

$$E(S_n) = E\left(\sum_{i=1}^{n} X_i\right) = 0$$

$$\text{Var}(X_n) = \text{Var}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \sum_{i=1}^{n} \text{Var}\, X_i + 2 \sum_{i<j=1}^{n} \text{Cov}\ (X_i, X_j)$$

$$= n + 2\,.\,(n-1)\,\rho$$ [On using (*)]

Since $X_1, X_2, ..., X_n$ are jointly normal,

$$S_n = \sum_{i=1}^{n} X_i \sim N(0, \sigma^2) \text{ where } \sigma^2 = n + 2(n-1)\,\rho > 0\,;$$

Taking $Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{S_n}{n}$, we have

$$E\left[\frac{Y_n^2}{1+Y_n^2}\right] = E\left[\frac{S_n^2}{n^2+S_n^2}\right]$$

$$= \int_{-\infty}^{\infty} \left(\frac{x^2}{n^2+x^2}\right) \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2/2\sigma^2}\, dx$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \frac{\sigma^2 y^2}{n^2+\sigma^2 y^2} \cdot e^{-y^2/2}\, dy \; ; \qquad \left(\frac{x}{\sigma} = y\right)$$

$$= \frac{2}{\sqrt{2\pi}} \{n + 2(n-1)\rho\} \cdot \int_0^{\infty} \frac{y^2}{n^2 + y^2 \{n + 2(n-1)\rho\}} \cdot e^{-y^2/2}\, dy$$

$$\leq \frac{n + 2(n-1)\rho}{n^2} \cdot \frac{2}{\sqrt{2\pi}} \int_0^{\infty} y^2 \cdot e^{-y^2/2}\, dy$$

$$\to 0 \text{ as } n \to \infty$$

$$\therefore \qquad \lim_{n \to \infty} E\left|\frac{Y_n^2}{1+Y_n^2}\right| \to 0$$

Hence by Theorem 6·33, WLLN holds for $\{X_n\}$, *i.e.*, $\frac{S_n}{n} \xrightarrow{p} 0$ as $n \to \infty$.

6·16. Borel-Cantelli Lemma. (*Zero-One Law*). Let A_1, A_2 ... be a sequence of events. Let A be the eent "that an infinite number of A_n occur. That is $\omega \in A$ if $\omega \in A_n$ for an infinite number of values of n (but not necessarily every n). But the set of such ω is precisely lim sup A_n, *i.e.*, $\overline{\lim}\, A_n$. Thus the event A, that an infinite number of A_n occur is just $\overline{\lim}\, A_n$. Sometimes we are interested in the probability that an infinite number of the events A_n occur. Often this question is answered by means of the Borel-Cantelli lemma or its converse.

Theorem 6·34. (*Borel-Cantelli Lemma*). *Let A_1, A_2, ... be a sequence of events on the probability space (S, B, P) and let $A = \overline{\lim}\, A_n$.*

if $\sum_{n=1}^{\infty} P(A_n) < \infty$, *then* $P(A) = 0$.

In other words, this states that if $\Sigma\, P(A_n)$ converges then with probability one, only a finite number of $A_1, A_2, \ldots$ can occur.

Proof. Since $\quad A = \overline{\lim}\, A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$

we have $A \subset \bigcup_{m=n}^{\infty} A_m$, for every n.

Thus for each n,

$$P(A) \le \sum_{m=n}^{\infty} P(A_m)$$

Since $\sum_{n=1}^{\infty} P(A_n)$ is convergent (given), $\sum_{m=n}^{\infty} P(A_m)$, being the remainder term of a convergent series, tends to zero as $n \to \infty$.

$$\therefore \qquad P(A) \le \sum_{m=n}^{\infty} P(A_m) \to 0 \text{ as } n \to \infty.$$

Thus $P(A) = 0$ as required.

The result just proved does not require events $A_1, A_2, \ldots$ considered to be independent. For the converse result it is necessary to make this further assumption.

Theorem 6·35. Borel-Cantelli Lemma (Converse). *Let* $A_1, A_2, \ldots$ *be independent events on* $(S, \mathbf{B}, \mathbf{P})$, A equal to $\overline{\lim}\ A_n$.

If $\sum_{n=1}^{\infty} P(A_n) = \infty$, *then* $P(A) = 1$.

Proof. Writing, in usual notation, $\overline{A_n}$ for the complement $S - A_n$ of A_n, we have for any m, n $(m > n)$.

$$\bigcap_{k=n}^{\infty} \overline{A_k} \subset \bigcap_{k=n}^{m} \overline{A_k}$$

$$P\left(\bigcap_{k=n}^{\infty} \overline{A_k}\right) \le P\left(\bigcap_{k=n}^{m} \overline{A_k}\right)$$

$$= \prod_{k=n}^{m} P(\overline{A_k}),$$

because of the fact that if $(A_n, A_{n+1}, \ldots, A_m)$ are independent events, so are $(\overline{A_n}, \overline{A_{n+1}}, \ldots, \overline{A_m})$.

Hence
$$P\left(\bigcap_{k=n}^{\infty} \overline{A_k}\right) \le \prod_{k=n}^{m} [1 - P(A_k)]$$

$$\le \prod_{k=n}^{m} e^{-P(A_k)}$$

$$[\because 1 - x \le e^{-x} \text{ for } x \ge 0]$$

$$= e^{-\prod_{k=n}^{m} P(A_k)} \qquad \forall\ m$$

Since
$$\sum_{k=1}^{\infty} P(A_k) = \infty, \quad \sum_{k=n}^{m} P(A_k) \to \infty \text{ as } m \to \infty$$

and hence $e^{-\sum_{k=n}^{m} P(A_k)}\ \forall\ m \to 0$ as $m \to \infty$.

$$\therefore \qquad P\left(\bigcap_{k=n}^{\infty} \overline{A_k}\right) = 0 \qquad \ldots(*)$$

But $$A = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

$\therefore$ $$\overline{A} = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \overline{A_k} \qquad \text{(De-Morgan's Law)}$$

$\Rightarrow$ $$P(\overline{A}) \le \sum_{n=1}^{\infty} P\left(\bigcap_{k=n}^{\infty} \overline{A_k}\right) = 0 \qquad \text{[From (*)]}$$

Hence $$P(A) = 1 - P(\overline{A}) = 1, \text{ as required.}$$

If $A_1, A_2, \ldots$ are independent events it follows from Theorems 6·34 and 6·35 that the probability that an infinite number of them occur is either zero (when $\sum_{n=1}^{\infty} P(A_n) < \infty$) or one [when $\sum_{n=1}^{\infty} P(A_n) = \infty$]. This statement is a special case of so-called "*Zero one law*" which we now state.

Theroem 6·36. (*Zero One Law*): *If $A_1, A_2, \ldots$ are independent and if E belongs to the σ-field generated by the class $(A_n, A_{n+1}, \ldots)$ for every n, then P (E) is zero or one.*

Example 6·62. *What is the probability that in a sequence of Bernoulli trials with probability of success p for each trial, the pattern SFS appears infinitely often ?*

Solution. Let A_k be the event that the trial number k, $K + 1$, $k + 2$ produce the sequence SFS ($k = 0, 1, 2, \ldots$). The events A_k are not mutually independent but the sequence $A_1, A_4, A_7, A_{10}, \ldots$ contains only mutually independent events (since no two depend on the outcome of the same trials).

$$p_k = P(A_k) = P(SFS) = p^2 q, \qquad (q = 1 - p)$$

is independent of k, and hence the series

$$p_1 + p_4 + p_7 + \ldots, \text{ diverges}$$

Hence by B.C.T. (converse) the pattern SFS appears infinitely often with probability one.

Example 6·63. *A bag contains one black ball and white balls. A ball is drawn at random. If a white ball is drawn, it is returned to the bag together with an additional white ball. If the black ball is drawn, it alone is returned to the bag.*

Let A_n denote the event that the black ball is not drawn in the first n trials. Discuss the converse to Borel-Cantelli Lemma with reference to events A_n.

Solution. A_n = The event that blackball is not drawn in the first n trials. ...(*)

= The event that each of the first n trials resulted in the draw of a white ball.

$\Rightarrow$ $$P(A_n) = P(E_1 \cap E_2 \cap \ldots \cap E_n),$$

where E_i is the event of drawing a white ball in the ith trial.

$\therefore$ $$P(A_n) = P(E_1)\, P(E_2 \mid E_1) \ldots P(E_3 \mid E_1 \cap E_2) \ldots P(E_n \mid E_1 \cap E_2 \ldots \cap E_{n-1})$$

$$= \frac{m}{m+1} \times \frac{m+1}{m+2} \times \ldots \times \frac{m+n-1}{m+n} = \frac{m}{m+n}$$

(Since if first ball drawn is white it is returned together with an additional white ball, *i.e.*, for the seconod draw the box contains 1 *b*, *m* + 1 *W* balls and

$$\therefore \qquad P(E_2 \mid E_1) = \frac{m+1}{m+2}, \text{ and so on.}$$

$$\sum_{n=1}^{\infty} PA(A_n) = \sum_{n=1}^{\infty} \frac{m}{m+n} = m \sum_{n=1}^{\infty} \frac{1}{m+n}$$

$$= m \left[\frac{1}{m+1} + \frac{1}{m+2} + \frac{1}{m+3} + \ldots \right]$$

$$= m \left[\sum_{n=1}^{\infty} \frac{1}{n} - \left(\sum_{n=1}^{m} \frac{1}{n} \right) \right] \qquad \ldots(**)$$

But $\sum_{n=1}^{\infty} \frac{1}{n}$ is divergent, $\left(\because \sum_{n=1}^{\infty} \frac{1}{n^p} \text{ is convergent, iff } p > 1 \right)$

and $\sum_{n=1}^{m} \frac{1}{n}$ is finite,

$\therefore$ R.H.S. of (**) is divergent.

Hence $\sum_{n=1}^{\infty} P(A_n) = \infty$

From the definition of A_n in (*) it is obvious that $A_n \downarrow$.

$$\therefore \qquad A = \overline{\lim_n}\, A_n = \limsup_n A_n = \varphi$$

$$\therefore \qquad P(A) = P(\varphi) = 0$$

This result is *inconsistent* with converse of Borel-Cantelli Lemma, the reason being that *the events* A_n $(n = 1, 2, \ldots)$ *considered here are not independent,*

$$\because \quad P(A_i \cap A_j) = P(A_j) = \frac{m}{m+i} \neq P(A_i)\, P(A_j),$$

since for $(j > i)\, A_j \subset A_i$ as $A_n \downarrow$

EXERCISE 6 (d)

1. State and prove Chebychev's inequality.
2. (a) A random variable X has a mean value of 5 and variance of 3.
   (*i*) What is the least value of Prob $[\,|X - 5| < 3]$?
   (*ii*) What value of h guarantees that Prob $[\,|X - 5| < h] \geq 0{\cdot}99$?
   (*iii*) What is the least value of Prob $(\,|X - 5| < 7{\cdot}5)$?

(*b*) A random variable X takes the values -1, 1, 3, 5 with associated probabilities 1/6, 1/6, 1/6 and 1/2. Find by direct computation $P(|X-3| \geq 1)$. Find an upper bound to this probability by applying Chebychev's inequality.

(*c*) If X denote the sum of the numbers obtained when two dice are thrown, use Chebychev's inequality to obtain an upper bound for $P\{|X-7| > 4\}$. Compare this with the actualy probability.

3. (*a*) An unbaised coin is tossed 100 times. Show that the probability that the number of heads will be between 30 and 70 is greater than 0·93.

(*b*) Within what limits will the number of heads lie, with 95 p.c. probability, in 1000 tosses of a coin which is practically unbised?

(*c*) A symmetric die is thrown 720 times. Use Chebychev's inequality to find the lower bound for the probability of getting 100 to 140 sixes.

(*d*) Use Chebychev's inequality to determine how many times a fair coin must be tossed in order that the probability will be at least 0·95 that the ratio of the number of heads to the number of tosses will be between 0·45 and 0·55.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

4. (*a*) If you wish to estimate the proportion of engineers and scientists who have studied probability theory and you wish your estimate to be correct within 2% with probability 0·95 or more, how large a sample would you take (*i*) if you have no idea what the true proportion is, (*ii*) if you are confident that the true proportion is less than 0·2? **[Burdwan Univ. (B.Sc. (Hons.), 1992]**

Hint. (*i*) $\varepsilon = 2\%$ or 0·02

$$P[|f_n - p| < 0{\cdot}02] \geq 1 - \frac{1}{4n\,(0{\cdot}02)^2} = 0{\cdot}95$$

$$\therefore \quad 0{\cdot}05 = \frac{1}{4n\,(0{\cdot}02)^2} \quad \Rightarrow \quad n = 12{,}500$$

(*ii*) $p < 0{\cdot}2,\ P[|f_n - p| \leq \varepsilon] \geq 1 - \dfrac{p(1-p)}{n\,\varepsilon^2}$

Now $p(1-p) < 0{\cdot}16$, therefore $1 - \dfrac{0{\cdot}16}{n\,\varepsilon^2} = 0{\cdot}95$

Hence $n = 50 \times 50 \times 20 \times {\cdot}16 = 8000$

(*b*) Let the sample mean of a random variable X be $\bar{X}$ s.d.s. Then if at lease 99 per cent of the values of X fall within K standard deviations from the mean, find K.

5. (*a*) If X is a r.v. such that $E(X) = 3$ and $E(X^2) = 13$, use Chebychev's inequality to determine a lower bound for $P(-2 < X < 8)$.

[Delhi Univ. B.Sc. (Maths Hons.), 1990]

Hint. $\mu_x = 3,\ \sigma_x^2 = 4 \ \Rightarrow \ \sigma_x = 2$. Chebychev's inequality gives

$P[|X-3| < 2k] \geq 1 - 1/k^2 \ \Rightarrow \ P(3 - 2k < X < 3 + 2k) \geq 1 - 1/k^2$

Now taking $k = 2{\cdot}5$, we get $P(-2 < X < 8) \geq 21/25$.

(*b*) State and prove Chebychev's inequality. Use it to prove that in 2000 throws with a coin the probability that the number of heads lies between 900 and 1100 is at least 19/20. **[Delhi Univ. B.Sc. (Maths Hons.), 1989]**

6. (*a*) A random variable X has the density function e^{-x} for $x \geq 0$. Show that Chebychev's inequality gives $P[|X-1|>2]<1/4$ and show that the actual probability is e^{-3}.

(*b*) Let X have the p.d.f.:

$$f(x) = \frac{1}{2\sqrt{3}}, -\sqrt{3} < x < \sqrt{3}$$
$$= 0, \text{ elsewhere}.$$

Find the actual probability $P\left[\,|X-\mu| \geq \frac{3}{2}\sigma\,\right]$ and compare it with the upper bound obtained by Chebychev's inequality.

7. If X has the distribution with p.d.f.

$$f(x) = e^{-x},\ 0 \leq x < \infty,$$

use Chebychev's inequality to obtain a lower bound to the probability of the inequality $-1 \leq X \leq 3$, and compare it with actual value.

8. Explain the concept of "convergence in probability".

If $X_1, X_2, \ldots, X_n$ by r.v.s. with means $\mu_1, \mu_2, \ldots, \mu_n$ and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_n$ and if $\sigma_n \to 0$ as $n \to \infty$, show that $\bar{X}_n - \bar{\mu}_n$ converges to zero stochastically.

Hence show that if m is the number of successes in n independent trials, the probability of success at the ith trial being p_i then m/n converges in probability to $(p_1 + p_2 + \ldots + p_n)/n$.

9. (*a*) If X_n takes the values 1 and 0 with corresponding probabilities p_n and $1-p_n$, examine whether the weak law of large numbers can be applied to the sequence $\{X_n\}$ where the variables X_n, $n = 1, 2, 3, \ldots$ are independent.

(*b*) $\{X_i\}$, $i = 1, 2, \ldots$ is a sequence of independent random variables with expected value of X_i equal to m_i and variance of X_i, equal to σ_i^2. If $\frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2$ tends to zero as n tends to infinity, show that the weak law of large numbers holds good to the sequence. **[Bombay Univ. B.Sc. (Stat.), 1992]**

10. $\{X_k\}$, $k = 1, 2, \ldots$ is a sequence of independent random variables each taking the values $-1, 0, 1$. Given that

$$P(X_k = 1) = \frac{1}{k} = P(X_k = -1),\ P(X_k = 0) = 1 - \frac{2}{k}.$$

Examine if the law of large numbers holds for this sequence.

11. (*a*) Derive Chebychev's inequality and show how it leads to the weak law of large numbers. Mention some important particular cases wherein the weak law of large numbers holds good.

(b) State and prove the weak law of large numbers. Deduce as a corollary Bernoulli theorem and comment on its applications.

(c) Examine whether the weak law of large numbers holds good for the sequence X_n of independent random variables where

$$P\left(X_n = \frac{1}{\sqrt{n}}\right) = \frac{2}{3},\ P\left(X_n = -\frac{1}{\sqrt{n}}\right) = \frac{1}{3}$$

(d) $\{X_n\}$ is a sequence of independent random variables such that

$$P\left(X_n = \frac{1}{\sqrt{n}}\right) = p_n\ ,\ P\left(X_n = 1 + \frac{1}{\sqrt{n}}\right) = 1 - p_n$$

Examine whether the weak law of large numbers is applicable to the sequence $\{X_n\}$

(e) If X is a random variable and $E(X^2) < \infty$, then prove that

$$P\{\,|X| \geq a\,\} \leq \frac{1}{a^2} E(X^2)\ ,\ \text{for all } a > 0\ .$$

Use Chebyshev's inequality to show that for $n > 36$, the probability that in n throws of a fair die, the number of sixes lies between $\frac{1}{6}n - \sqrt{n}$ and $\frac{1}{6}n + \sqrt{n}$ is at least $31/36$. [Calcutta Univ. B.Sc. (Maths Hons.), 1991]

12. Let $\{X_n\}$ be a sequence of mutually independent random variables such that

that $\qquad X_n = \pm 1$ with probability $\dfrac{1 - 2^{-n}}{2}$

and $\qquad X_n = \pm 2^{-n}$ with probability 2^{-n-1}

Examine whether the weak law of large numbers can be applied to the sequence $\{X_n\}$.

13. Examine whether the law of large numbers holds for the sequence $\{X_k\}$ of independent random variables defined by $P(X_k = \pm k^{-1/2}) = \frac{1}{2}$.

14. (a) State Khinchin's theorem.

(b) Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed r.v.'s, each uniform on $[0, 1]$. For the geometric mean

$$G_n = (X_1 X_2 \ldots X_n)^{1/n}$$

show that $G_n \xrightarrow{p} c$ for some finite number c. Find c.

Hint. $X \sim U[0, 1]$, let $Y = -\log X$;

Then $F_Y(y) = 1 - e^{-y}$; $\Rightarrow$ $f_Y(y) = e^{-y}$; $y \geq 0$.

$\therefore\ Y_i = -\log X_i$, $(i = 1, 2, \ldots, n)$ are i.i.d. r.v.'s. with $E(Y_i) = 1$.

By Khinchin's theorem

$$\sum_{i=1}^{n} Y_n/n = -\left(\sum_{i=1}^{n} \log X_i/n\right) = -\log G_n \xrightarrow{p} E\,Y_i = 1 \quad \Rightarrow \quad G_n \xrightarrow{p} e^{-1} = c\ .$$

(c) Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with common p.d.f.

$$f(x) = \frac{1+\delta}{x^{2+\delta}}, x \geq 1, \ \delta > 0$$
$$= 0, \ x < 1$$

Discuss if WLLN holds for the sequence $\{X_n\}$.

Hint. $E X_i = (1 + \delta)/\delta < \infty$ (finite). Hence by Khinchin's theorem

$$S_n/n = \Sigma X_i/n \xrightarrow{p} (1+\delta)/\delta \text{ as } n \to \infty.$$

15. Let $\{X_n\}$ be any sequence of r.v.'s Write $Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Prove that a necessary and sufficient condition for the sequence $\{X_n\}$ to satisfy the weak law of large numbers is that

$$E\left[\frac{Y_n^2}{1+Y_n^2}\right] \to \text{ as } n \to \infty.$$

Hint. See Remark to Theorem 6·33.

16. State and prove Weak Law of Large Numbers. Determine whether it holds for the following sequence of independent random variables:

$$P(X_n = +1) = (1 - 2^{-n})/2 = P(X_n = -1)$$

[Delhi Univ. B.Sc. (Maths Hons.), 1989]

17. Let $X_1, X_2, \ldots$ be i.i.d. standard Cauchy variates. Show that the WLLN does not hold for the sequence $\{X_n\}$.

Hint. Use Theorem 6·33.

$$E\left(\frac{Y_n^2}{1+Y_n^2}\right) = E\left(\frac{S_n^2}{n^2+S_n^2}\right) = E\left(\frac{(S_n/n)^2}{1+(S_n/n)^2}\right)$$

$$= \int_{-\infty}^{\infty} \frac{x^2}{1+x^2} \cdot \frac{1}{\pi} \frac{1}{1+x^2} dx$$

$\left[\because \frac{S_n}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}\right.$ is also a standard Cauchy variate. See Remark 4, § 8·9·1]

$$= \frac{2}{\pi} \int_0^{\infty} \sin^2 \theta \, d\theta = \frac{1}{2} \qquad (x = \tan \theta)$$

$$\Rightarrow \lim_{n \to \infty} E\left[\frac{Y_n^2}{1+Y_n^2}\right] \not\to 0 \Rightarrow \text{WLLN does not hold for } \{X_n\}.$$

18. (*a*) Examine if the WLLN holds for the sequence $\{X_n\}$ of i.i.d. r.v.'s with

$$P\left[X_i = (-1)^{k-1} . k\right] = \frac{6}{\pi^2 k^2}; \ k = 1, 2, 3, \ldots, i = 1, 2, 3, \ldots$$

[Delhi Univ. B.Sc. (Maths Hons.), 1990]

Hint. $E(X_i) = \sum_{k=1}^{\infty} (-1)^{k-1} \cdot k \cdot \dfrac{6}{\pi^2 k^2} = \dfrac{6}{\pi^2} \sum_{k=1}^{\infty} (-1)^{k-1} \cdot (1/k)$

$$= \frac{6}{\pi^2}\left[1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \dots\right]$$

$$= \frac{6}{\pi^2} \cdot \log_e 2$$

[$\because$ The series in bracket is convergent by Leibnitz test for alternating series].

Hence by Khinchin's theorem, WLLN holds for the sequence $\{X_i\}$ of i.i.d. r.v.'s.

(*b*) The r.v.'s $X_1, X_2, \dots, X_n$ have equal expectations and finite variation. Is the weak law of large numbers applicable to this sequence if all the co-variances σ_{ij} are negative? **[Delhi Univ. B.Sc. (Maths Hons.), 1987]**

Hint. $\dfrac{B_n}{n^2} = \dfrac{\text{Var}(X_1 + X_2 + \dots + X_n)}{n^2} = \dfrac{1}{n^2}\left[\sum_{i=1}^{n} \sigma_i^2 + 2 \sum_{i<j=1}^{n} \sigma_{ij}\right]$

$$< \frac{1}{n^2}\left(\sum_{i=1}^{n} \sigma_i^2\right) \to 0 \text{ as } n \to \infty \qquad (\because \sigma_i^2 \text{ are finite})$$

Hence WLLN holds.

19. State and prove Borel Cantelli Lemma.

In a sequence of Bernoulli trials let A_n be the event that a run of n consecutive successes occurs between the 2^nth and 2^{n+1}th trails. Show that if $p \geq \frac{1}{2}$, there is probability one that infinitely many A_n occur; if $p < \frac{1}{2}$, then with probability one only finitely many A_n occur.

20. Let $X_1, X_2, \dots$ be independent r.v.'s and $S_n = \sum_{k=1}^{n} X_k$. If $\sum_{n=1}^{\infty} \sigma^2_{X_n}$ converges, prove that the series $\Sigma (X_n - E X_n)$ converges in probability.

If $\dfrac{1}{b_n^2} \sum_{k=1}^{n} \sigma^2_{X_k} \to 0$, then prove that

$$\frac{S_n - ES_n}{b_n} \xrightarrow{p} 0.$$

Deduce Chebychev's inequality.

6·17. Probability Generating Function

Definition. *If $a_0, a_1, a_2, \dots$ is a sequence of real numbers and if*

$$A(s) = a_0 + a_1 s + a_2 s^2 + \dots = \sum_{i=0}^{\infty} a_i s^i \qquad \dots(6\cdot84)$$

converges in some interval $-s_0 < s < s_0$, when the sequence is infinite then the function $A(s)$ is known as the generating function of the sequence $\{a_i\}$.

The variable s has no significance of its own and is introduced to identify a_i as the co-efficient of s^i in the expansion of $A(s)$. If the sequence $\{a_i\}$ is bounded, then the comparison with the geometric series shows that $A(s)$ converges at least for $|s| < 1$.

In the particular case when a_i is the probability that an integral valued discrete variable X takes the value i,

i.e., $a_i = p_i = P(X = i)$; $i = 0, 1, 2, \ldots$ with $\Sigma p_i = 1$, then the probability generating function, abbreviated as p.g.f., of r.v. X is defined as :

$$P(s) = E(s^X) = \sum_{x=0}^{\infty} s^x \cdot p_x \qquad \ldots(6{\cdot}85)$$

Remarks. 1. Obviously we have $P(1) = \sum_x p_x = 1$.

Thus a function $P(s)$ defined in (6·85) is a p.g.f. iff $p_x \geq 0 \ \forall \ x$ and $\sum_x p_x = 1$

2. *Relation between p.g.f. and m.g.f.*

Taking $s = e^t$ in (6·85), we get

$$P(e^t) = E(e^{tX}) = M_X(t). \qquad \ldots(6{\cdot}86)$$

i.e., from p.g.f. we can obtain m.g.f. on replacing s by e^t.

3. *Bivariate probability generating function.* The joint p.g.f. of two random variables X_1 and X_2 is a function of two variables s_1 and s_2 defined by :

$$P_{X_1, X_2}(s_1, s_2) = E(s_1^{X_1} \cdot s_2^{X_2}) = \sum_{x_1} \sum_{x_2} s_1^{x_1} s_2^{x_2} \cdot p(x_1, x_2) \qquad \ldots(6{\cdot}87)$$

Marginal p.g.f.'s can be obtained from (6·87) as given below.

$$P_{X_1}(s_1) = E(s_1^{X_1}) = P_{X_1, X_2}(s_1, 1) \ ; \ P_{X_2}(s_2) = E(s_2^{X_2}) = P_{X_1, X_2}(1, s_2) \qquad \ldots(6{\cdot}88)$$

4. Two r.v.'s are *independent* if and only if :

$$P_{X_1, X_2}(s_1, s_2) = P_{X_1}(s_1) \cdot P_{X_2}(s_2). \qquad \ldots(6{\cdot}89)$$

The above concepts can be generalised to n random variables

Theorem 6·37. *If X is a random variable which assumes only integral values with probability distribution*

$$P(X = k) = p_k \ ; \ k = 0, 1, 2, \ldots \ \textit{and} \ P(X > k) = q_k, \ k \geq 0$$

so that $q_k = p_{k+1} + p_{k+2} + \ldots = 1 - \sum_{i=0}^{k} p_i$ *and two generating functions are*

$$P(s) = p_0 + p_1 s + p_2 s^2 + \ldots$$

$$Q(s) = q_0 + q_1 s + q_2 s^2 + \ldots$$

then for $\quad -1 < s < 1, \ Q(s) = \dfrac{1 - P(s)}{1 - s} \qquad \ldots(6{\cdot}90)$

Proof. We have

$$q_{k-1} - q_k = p_k\,,\quad k \ge 1$$

$$\sum_{k=1}^{\infty} q_{k-1} s^k - \sum_{k=1}^{\infty} q_k s^k = \sum_{k=1}^{\infty} p_k s^k$$

$\Rightarrow$ $$s\,Q(s) - Q(s) + q_0 = P(s) - p_0$$

$\Rightarrow$ $$Q(s) = \frac{(p_0 + q_0) - P(s)}{(1-s)}$$

But $\quad p_0 + q_0 = p_0 + p_1 + p_2 + \dots = 1$

Hence the theorem.

Theorem 6·38. *For a random variable X, which assumes only integral values, the expectation $E(X)$ can be calculated either from the probability distribution $P(X = i) = p_i$ or in terms of*

$$q_k = p_{k+1} + p_{k+2} + \dots$$

Thus $$E(X) = \sum_{i=1}^{\infty} i p_i = \sum_{k=0}^{\infty} q_k$$

In terms of the gererating functions

$$E(X) = P'(1) = Q(1) \qquad \text{...(6·91)}$$

Proof. $P(s) = \sum_{k=0}^{\infty} p_k s^k$. If $E(X)$ exists, then

$$E(X) = \sum_{k=1}^{\infty} k p_k$$

$$P'(s) = \sum_{k=1}^{\infty} k p_k s^{k-1} \Rightarrow P'(1) = \sum_{k=1}^{\infty} k p_k$$

$\therefore$ $$E(X) = P'(1)$$

We know that

$$Q(s)[1-s] = 1 - P(s)$$

Differentiating both sides w.r.t. s, we get

$$Q'(s)[1-s] - Q(s) = -P'(s) \qquad \text{...(*)}$$

$\therefore$ $$Q(1) = P'(1)$$

Hence $$E(X) = P'(1) = Q(1)$$

Theorem 6·39. *If $E(X^2) = \Sigma k^2 p_k$ exists, then*

$$E(X^2) = P''(1) + P'(1) = 2Q'(1) + Q(1)$$

and hence $V(X) = 2Q'(1) + Q(1) - \{Q(1)\}^2 = P''(1) + P'(1) - \{P'(1)\}^2$ ...(6·92)

Proof. $P(s) = \sum_{k=0}^{\infty} p_k s^k$, $P'(s) = \Sigma k p_k s^{k-1}$

$$s\,P'(s) = \sum k\,p_k\,s^k. \text{ Differentiating again, we get}$$

$$P'(s) + sP''(s) = \sum k^2 p_k\,s^{k-1}$$

$$\therefore \quad P'(1) + P''(1) = \sum_k k^2 p_k = E(X^2) \qquad \ldots(**)$$

$$\therefore \quad Q''(s)[1-s] - 2Q'(s) = -P''(s) \qquad \text{[Differentiating (*) again]}$$

Putting $s = 1$, we get

$$2Q'(1) = P''(1). \text{ Substituting in (**), we get}$$

$$E(X^2) = P'(1) + P''(1) = Q(1) + 2Q'(1)$$

$$\therefore \quad \text{Var}(X) = E(X^2) - \{E(X)\}^2 = P''(1) + P'(1) - \{P'(1)\}^2$$

$$= 2Q'(1) + Q(1) - \{Q(1)\}^2$$

6·17·1. Probability Generating Function for the sum of independent variables (Convolutions). If X and Y are non-negative independent, integral valued discrete random variables with respective probability generating functions,

$$P(s) = \sum_{k=0}^{\infty} p_k\,s^k, \quad p_k = P(X = k)$$

$$R(s) = \sum_{k=0}^{\infty} r_k\,s^k, \quad r_k = P(Y = k),$$

it is possible to deduce the probability generating function for the variable $Z = X + Y$, which is also clearly integral valued, in terms of $P(s)$ and $Q(s)$.

Let w_k denote $P(Z = k)$. The event $Z = k$ is the union of the following mutually exclusive events,

$$(X = 0 \cap Y = k), (X = 1 \cap Y = k-1), (X = 2 \cap Y = k-2), \ldots, (X = k \cap Y = 0)$$

and

since the variables X and Y are independent, each joint probability is the product of the appropriate individual probabilities. Therefore the distribution $w_k = P(Z = k)$ is given by

$$w_k = p_0 r_k + p_1 r_{k-1} + p_2 r_{k-2} + \ldots + p_k r_0 \text{ for all integral } k \geq 0$$

The new sequence of probabilities $\{w_k\}$ defined in terms of the sequences $\{p_k\}$ and $\{r_k\}$ is called the convolution of these sequences and is denoted by

$$\{w_k\} = \{p_k\} * \{r_k\} \qquad \ldots(6·93)$$

Theorem 6·40.

$$\mu'_{(r)} = E\left[X(X-1)\ldots(X-r+1)\right] = \left[\frac{\partial^r}{\partial s^r} P(s)\right]_{s=1}$$

Proof. Differentiating (6·85) partially r times w.r.t. s, we get

$$\frac{\partial^r P(s)}{\partial s^r} = \sum_x x(x-1)(x-2)\ldots(x-r+1)\,s^{x-r}\,p_x$$

$$\left[\frac{\partial^r P(s)}{\partial s^r}\right]_{s=1} = \sum_x x(x-1)(x-2)\dots(x-r+1)\,p_x = \mu'_{(r)}$$

Theorem 6·41. *If* $\{p_k\}$ *and* $\{r_k\}$ *are the sequences with the generating functions* $P(s), R(s)$ *and* $\{w_k\}$ *is their convolution, then* $W(s) = P(s)R(s)$, *where* $W(s) = \Sigma\, w_k s^k$ *is the generating function of the sum* $X + Y$.

Proof. Since the co-efficient of s^k in the product $P(s)R(s)$ is

$$p_0 r_k + p_1 r_{k-1} + \dots + p_{k-1} r_1 + p_k r_0 = w_k,$$

it follows that the probability generating function for Z, namely,

$W(S) = \sum_{k=0}^{\infty} w_k s^k$ is equal to $P(s)R(s)$.

Cor. If $X_1, X_2, \dots, X_n$ are independent integral–valued discrete variables with respective probability generating functions $P_1(s), P_2(s), \dots, P_n(s)$ and if $Z = X_1 + X_2 + \dots + X_n$, the probability generating function for Z is given by

$$P_Z(s) = \prod_{i=1}^{n} P_i(s)$$

In particular, when $X_1, X_2, \dots, X_n$ all have a common distribution and hence common probability generating function $P(s)$, we have

$$P_Z(s) = [P(s)]^n$$

Example 6·64. *Can* $P(s) = 2/(1+s)$ *be the p.g.f. of a r.v.* X? *Give reasons.*

Solution. We have $P(1) = 2/2 = 1$.

Also $\quad P(s) = \sum_{r=0}^{\infty} p_r s^r = 2\,(1+s)^{-1}$

$$= 2\,(1 - s + s^2 - s^3 + \dots)$$

$$= 2\sum_{r=0}^{\infty} (-1)^r . s^r \qquad \dots(*)$$

$\Rightarrow \qquad p_r = 2(-1)^r$

Hence $p^{2k+1} < 0$, *i.e.*, $p_1, p_3, p_5, \dots$ are negative.

Since some co-efficient in $(*)$ are negative, $P(s)$ cannot be the p.g.f. of a r.v. X.

Example 6·65. *If* $P(s)$ *is the probability generating function for* X, *find the generating function for* $(X - a)/b$.

Solution. $\quad P(s) = E(s^X)$

P.G.F. for $\dfrac{X-a}{b} = E\left(s^{(x-a)/b}\right) = e^{-a/b} . E\left(s^{x/b}\right)$

$$= s^{-a/b} \cdot E\,[(s^{1/b})^x] = s^{-a/b}\,P\,(s^{1/b})$$

Example 6·66. *Let X be a random variable with generating function $P\,(s)$. Find the generating function of (a) $X+1$ (b) $2X$.*

Solution. *(a)* $P\,(s) = \sum_{k=0}^{\infty} p_k\, s^k = E\,(s^X)$

$\therefore$ P.G.F. of $X + 1 = E\,(s^{X+1}) = s\,.E\,(s^X) = s\,.\,P\,(s)$

(b) P.G.F. of $2X = E\,(s^{2X}) = E\,[(s^2)^X] = P\,(s^2)$

Example 6·67. *Find the generating function of (a) $P\,(X \le n)$, (b) $P\,(X < n)$, and (c) $P\,(X = 2n)$.* **[Delhi Univ. M.Sc. (O.R.), 1989]**

Solution. *(a)* Let X be an integral valued random variable with the probability distribution

$$P\,(X = n) = p_n \text{ and } P\,(X \le n) = q_n$$

so that $q_n = p_0 + p_1 + p_2 + \ldots + p_n\,;\; n = 0, 1, 2, \ldots$

$$\therefore \qquad q_n - q_{n-1} = p_n\,,\quad n \ge 1$$

$$\Rightarrow \quad \sum_{n=1}^{\infty} q_n s^n - \sum_{n=1}^{\infty} q_{n-1} s^n = \sum_{n=1}^{\infty} p_n s^n$$

$$\Rightarrow \quad Q(s) - q_0 - s\,Q(s) = P(s) - p_0$$

$$\Rightarrow \quad Q\,(s) = \frac{P\,(s) + q_0 - p_0}{1-s} = \frac{P\,(s)}{1-s} \qquad [\because q_0 = p_0]$$

(b) Let $q_n = P\,(X < n) = p_0 + p_1 + \ldots + p_{n-1},\; q_n - q_{n-1} = p_{n-1}\,,\; n \ge 2$

$$\Rightarrow \quad \sum_{n=2}^{\infty} q_n s^n - \sum_{n=2}^{\infty} q_{n-1} s^n = \sum_{n=2}^{\infty} p_{n-1} s^n = s \sum_{n=1}^{\infty} p_n s^n$$

$$\Rightarrow \quad Q\,(s) - q_1 s - sQ(s) = sP(s) - sp_0 \qquad [\because q_0 = 0]$$

$$\Rightarrow \quad Q\,(s)\,[1 - s] = sP(s) - sp_0 + q_1 s \qquad [\because q_1 = p_0]$$

Hence $$Q\,(s) = \frac{sP(s)}{1-s}$$

(c) Let $P(X = 2n) = p_{2n}$

$$\therefore \quad Q\,(s) = \sum_{n=0}^{\infty} p_{2n}\, s^n = p_0 + p_2 s + p_4 s^2 + \ldots$$

$$2Q\,(s) = 2p_0 + 2p_2\, s + 2p_4\, s^2 + \ldots$$

$$= (p_0 + p_1 s^{1/2} + p_2 s + p_3 s^{3/2} + p_4 s^2 + \ldots) + (p_0 - p_1 s^{1/2} + p_2 s - p_3 s^{3/2} + \ldots)$$

$$= \sum_{k=0}^{\infty} p_k\,(s^{1/2})^k + \sum_{k=0}^{\infty} p_k\,(-s^{1/2})^k$$

$$= P\,(s^{1/2}) + P\,(-s^{1/2})$$

$$Q\,(s) = \frac{P\,(s^{1/2}) + P\,(s^{-1/2})}{2}$$

Example 6·68. *Let $\{X_k\}$ be mutually independent, each assuming the values $0, 1, 2, ..., a-1$ with probability $\frac{1}{a}$.*

Let $S_n = X_1 + X_2 + ... + X_n$. Show that the generating function of S_n is

$$P(s) = \left[\frac{1-s^a}{a(1-s)}\right]^n$$

and hence

$$P(S_n = j) = \frac{1}{a^n}\sum_{\nu=0}^{\infty}(-1)^{\nu+j+a\nu}\binom{n}{\nu}\binom{-n}{j-a\nu}$$

(Rajasthan Univ. M.Sc. 1992)

Solution. As the $\{X_k\}$ are mutually independent variables and each X_i assumes the same values $0, 1, 2, ..., a-1$, with the probability $\frac{1}{a}$, therefore each will have the same generating function and the generating function of S_n will be the nth convolution of generating function of X_1. Now

$$P_{X_1}(s) = \sum_{k=0}^{a-1} p_k s^k = \frac{1}{a}[s^0 + s^1 + ... + s^{a-1}] = \frac{1-s^a}{a(1-s)}$$

$$\therefore \quad P_{S_n}(s) = \left[\frac{1-s^a}{a(1-s)}\right]^n$$

Now the probability $S_n = j$ is the co-efficient of s^j in

$$\frac{1}{a^n}(1-s^a)^n(1-s)^{-n}$$

If we take $(\nu+1)$th term from $(1-s^a)^n$, then it will have the power of s equivalent to $s^{a\nu}$ and hence to get the power of s as j, we must take the term from $(1-s)^{-n}$ having the power of s as $j - a\nu$.

$\therefore$ Required probability

$$= \frac{1}{a^n}\sum_{\nu=0}^{\infty}\binom{n}{\nu}(-1)^{\nu}\binom{-n}{j-a\nu}(-1)^{j-a\nu}$$

$$= \frac{1}{a^n}\sum_{\nu=0}^{\infty}(-1)^{j+\nu-a\nu}\binom{n}{\nu}\binom{-n}{j-a\nu}$$

$$= \frac{1}{a^n}\sum_{\nu=0}^{\infty}(-1)^{j+\nu+a\nu}\binom{n}{\nu}\binom{-n}{j-a\nu}$$

$$\left[\because (-1)^{2a\nu} = 1 \Rightarrow (-1)^{a\nu} = (-1)^{-a\nu}\right]$$

Example 6·69. *A random variable X assumes the value $\lambda_1, \lambda_2, ...$ with probabilities $u_1, u_2, ...$, show that*

$$p_k = \frac{1}{k!}\sum_{j=0}^{\infty} u_j e^{-\lambda_j}(\lambda_j)^k; \quad \lambda_j > 0, \ \Sigma u_j = 1$$

is a probability distribution. Find its generating function and prove that its mean equals E (X) and variance equals V (X) + E (X).

Solution.

$$\sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} \left[\frac{1}{k!} \sum_{j=0}^{\infty} u_j e^{-\lambda_j} (\lambda_j)^k \right]$$

$$= \sum_{j=0}^{\infty} \left[u_j e^{-\lambda_j} \sum_{k=0}^{\infty} (\lambda_j)^k / k! \right] = \sum_{j=0}^{\infty} u_j e^{-\lambda_j} e^{\lambda_j}$$

$$= \sum_{j=0}^{\infty} u_j = 1$$

Hence p_k represents a prbability distribution.

Let $P(s)$ be the generating function of p_k, then

$$P(s) = \sum_{k=0}^{\infty} p_k s^k = \sum_{k=0}^{\infty} \left[\frac{1}{k!} \sum_{j=0}^{\infty} u_j e^{-\lambda_j} (\lambda_j)^k s^k \right]$$

(Fubini's Theorem)

$$= \sum_{j=0}^{\infty} \left[u_j e^{-\lambda_j} \sum_{k=0}^{\infty} (s\lambda_j)^k / k! \right]$$

$$= \sum_{j=0}^{\infty} u_j e^{-\lambda_j} e^{s\lambda_j} = \sum_{j=0}^{\infty} u_j e^{\lambda_j (s-1)}$$

Thus $$P(s) = \sum_{j=0}^{\infty} u_j \left[1 + \lambda_j (s-1) + \frac{\lambda_j^2 (s-1)^2}{2!} + \dots \right]$$

$$P'(1) = \sum_{j=0}^{\infty} u_j \lambda_j = E(X)$$

$$P''(s) = \sum_{j=0}^{\infty} u_j [\lambda_j^2 + \lambda_j^3 (s-1) + \dots]$$

$$\therefore \quad P''(1) = \sum_{j=0}^{\infty} \lambda_j^2 u_j = E(X^2)$$

$$V(p_k) = P''(1) + P'(1) - \{P'(1)\}^2 = E(X^2) + E(X) - \{E(X)\}^2$$
$$= E(X) + V(X)$$

EXERCISE 6 (e)

1. *(a)* Define the probability generating function (p.g.f.) of a random variable.

(b) X is a positive integral valued variable, such that $P(X = n) = p_n$, $n = 0, 1, 2, \dots$ Define the probability generating function $G(s)$ and the moment generating function $M(t)$ for X and show that, $M(t) = G(e^t)$. Hence or otherwise prove that

$$E(X) = G'(1), \text{ var}(X) = G''(1) + G'(1) - [G'(1)]^2$$

(c) If X and Y are non-negative integral valued independent random variables with $P(s)$ and $Q(s)$ as their probability generating functions, show that their sum $X + Y$ has the p.g.f. $P(s)\,Q(s)$.

2. A test of the strength of a wire consists of bending and unbending until it breaks. Considering bending and unbending as two operations, let X denote the random variable corresponding to the number of operations necessary to break the wire. If $P(X = r) = (1 - p)p^{r-1}$; $r = 1, 2, 3, \ldots$ and $0 < p < 1$, find the probability generating function of X.

3. Define the generating function $A(s)$ of the sequence $\{a_j\}$. Let a_j be the number of ways in which the score j can be obtained by throwing a die any number of times. Show that the g.f. of $\{a_j\}$ is $(1 - s - s^2 - s^3 - s^4 - s^5 - s^6)^{-1} - 1$.

4. Four tickets are drawn, one at a time with replacement, from a set of ten tickets numbered respectively 1, 2, 3, ..., 10, in such a way that at each draw each ticket is equally likely to be selected. What is the probability that the total of the numbers on the four drawn tickets is 20?

Hint. If X_i denotes the number on the ith ticket then, for $i = 1, 2, 3, 4$, we observe that X_i is an integral-valued variate with possible values 1, 2, 3, ..., 10, each having associated probability 1/10. Here each X_i has

$$\text{g.f.} = \frac{1}{10}s + \frac{1}{10}s^2 + \ldots + \frac{1}{10}s^{10} = \frac{1}{10}s\,(1 - s^{10})\,(1 - s)^{-1}$$

and, since the X_i's are independent, it follows that the total of the numbers on the drawn tickets

$$Z = X_1 + X_2 + X_3 + X_4$$

has probability generating function

$$\left\{\frac{1}{10}s\,(1 - s^{10})\,(1 - s)^{-1}\right\}^4 = \frac{1}{10^4}s^4\,(1 - s^{10})^4\,(1 - s)^{-4}$$

The required probability is the co-efficient of s^{16} in

$$\frac{1}{10^4}(1 - s^{10})^4\,(1 - s)^{-4} = \frac{63}{10{,}000}.$$

5. Find the generating functions of (a) $P(X \geq n)$, (b) $P(X > n + 1)$.

6. (a) Obtain the generating function of q_n, the probability that in n tosses of an ideal coin, no run of three heads occurs.

(b) Let X be a non-negative integral-valued random variable with probability generating function $P(s) = \sum_{n=0}^{\infty} p_n s^n$. After observing X, conduct X binomial trials with probability p of success and let Y denote the corresponding resulting number of successes.

Determine (i) the probability generating function of Y and, (ii) probability generating function of X given that $Y = X$.

7. In a sequence of Bernoulli trials, let U_n be the probability that the first combination SF occurs at trials number $(n-1)$ and n.

Find the generating function, mean and variance.

Hint. $U_2 = P(SF) = pq,\ U_3 = P(SSF) + P(FSF) = pq(p+q)$

$U_4 = (SSSF) + P(FSSF) + P(FFSF) = pq(p^2 + pq + q^2)$

In general $U_n = pq \sum_{k=0}^{n-2} p^k q^{n-2-k}$

$$= pq^{n-1}\left[1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2 + \dots + \left(\frac{p}{q}\right)^{n-2}\right]$$

$$= pq \cdot \frac{q^{n-1} - p^{n-1}}{q-p},$$

8. *(a)* In a sequence of Bernoulli trials, let U_n be the probability of an even number of successes. Prove the recursion formula

$$U_n = qU_{n-1} + (1 - U_{n-1})\,p$$

From this derive the generating function and hence the explicit formula for U_n.

(b) A series of independent Bernoulli trials is performed until an uninterrupted run of r successes is obtained for the first time where r is a given positive integer. Assuming that the probability of a success in any trial is $p = 1 - q$, show that the probability generating function of the number of trials is

$$F(s) = \frac{p^r s^r (1 - ps)}{1 - s + q p^r s^{r+1}}$$

ADDITIONAL EXERCISES ON CHPATER VI

1. *(a)* A horizontal line of length '5' units is divided into two parts. If the first part is of length X, find $E(X)$ and $E[X(5-X)]$.

(b) Show that

$$E(X-\mu)^3 = E(X^3) - 3\mu\sigma^2 - \mu^3$$

where μ and σ^2 are the mean and variance of X respectively.

2. *(a)* Two players A and B alternately roll a pair of fair dice. A wins if he gets six points before B gets seven points and B wins if he gets seven points before A gets six points. If A takes the first turn, find the probability that B wins and the expected number of trials for A to win.

(b) A box contains 2^n tickets among which ${}^nC_i$ tickets bear the numbers i $(i = 0, 1, 2, \dots, n)$. A group of m tickets is drawn. Let S denote the sum of their numbers. Find the expectation and variance of S.

Ans. $\frac{1}{2}mn,\ \frac{1}{4}mn - \{mn(m-1)/4(2^n-1)\}$

3. In an objective type examination, consisting of 50 questions, for each question there are four answers of which only one is correct. A candidate scores 1 if he picks up the correct answer and $-1/3$ otherwise. If a candidate makes only a random choice in respect of each of the 50 questions, find his expected score and the variance of his score.

4. *(a)* A florist, in order to satisfy the needs of a number of regular and sophisticated customers, stocks a highly perishable flower. A dozen flowers cost Rs. 3 and sell for Rs. 10. Any flowers not sold the day they are stocked are worthless. Demand in dozens of flowers is as follows :

Demand	0	1	2	3	4	5
Probability	0·1	0·2	0·3	0·2	0·1	0·1

(i) How many flowers should the florist stock daily in order to maximise the expected value of his net profit ?

(ii) Assuming that failure to satisfy any one customer's request will result in future lost profits amounting to Rs. 5·10 (goodwill cost), in addition to the lost profit on the immediate sale, how many flowers should the florist stock ?

(iii) What is the smallest goodwill cost of stocking five dozen flowers ?

Hint. For $i = 0, 1, 2, 3, 4, 5$, let X_i be the random variable giving the florist's net profit, when he decides to stock 'i' dozen flowers. Determine the probability function for each and the mean of each and pick up that 'i' for which it is maximum.

Ans. *(i)* 3 dozen, *(ii)* 4 dozen and *(iii)* Rs. 2

5. Consider a sequence of Bernoulli trials with a constant probability p of success in a single trial. Let X_k denote the number of failures following the $(k-1)$th and preceding the kth success, and let $S_r = \sum_{k=1}^{r} X_k$.

Derive the probability distribution of X_k. Hence derive the probability distribution of S_r. Find $E(S_r)$ and $\text{Var}(S_r)$.

6. In the simplest type of weather forecasting — "rain" or "no rain" in the next 24 hours — suppose the probability of raining is p $(> \frac{1}{2})$, and that a forecaster scores a point if his forecast proves correct and zero otherwise. In making n independent forecasts of this type, a forecaster, who has no genuine ability, predicts "rain" with probability λ and "no rain" with probability $(1-\lambda)$. Prove that the probability of the forecast being correct for any one day is

$$[1-p+(2p-1)\lambda]$$

Hence derive the expectation of the total score (S_n) of the forecaster for the n days, and show that this attains its maximum value for $\lambda = 1$. Also, prove that

$$\text{Var}(S_n) = n\,[p-(2p-1)\lambda]\,[1-p+(2p-1)\lambda]$$

and thereby deduce that, for fixed n, this variance is maximum for $\lambda = \frac{1}{2}$.

Hint. $P(X_i = 1) = 1 - P(X_i = 0) = p\lambda + q(1 - \lambda)$, $S_n = \sum_{i=1}^{n} X_i$,

and the X_i s being independent, the stated results follow.

7. In the simplest type of weather forecasting — rain or no rain in the next 24 hours — suppose the probability of raining is p $(> \frac{1}{2})$, and that a forecaster scores a point if his forecast proves correct and zero otherwise. In making n indiependent forecasts of this type, a forecaster who has no genuine ability decides to allocate at random r days to a "rain" forecast and the rest to "no rain". Find the expectation of his total score $\{S_n\}$ for the n days and show that this attains its maximum value for $r = n$. What is the variance of S_n ?

Hint. Let X_i be a random variable such that

$$X_i = 1 \text{ if forecast is correct for ith day}$$
$$= 0 \text{ if forecast is incorrect for ith day, } (i = 1, 2, ..., n)$$

Then $P(X_i = 1) = \dfrac{r}{n}$, $P(X_i = 0) = 1 - \dfrac{r}{n} = \dfrac{n-r}{n}$

$$\therefore \quad E(X_i) = p\left(\frac{r}{n}\right) + q\left(\frac{n-r}{n}\right) \text{ and } S_n = \sum_{i=1}^{n} X_i$$

But the X_i 's are correlated random variables, so that for $i \neq j$,

$$E(X_i X_j) = P(X_i = 1 \cap X_j = 1)$$
$$= p \cdot \frac{r}{n}\left[p\left(\frac{r-1}{n-1}\right) + q\left(\frac{n-r}{n-1}\right)\right]$$
$$+ q\left(\frac{n-r}{n}\right)\left[p\left(\frac{r}{n-1}\right) + q\left(\frac{n-r-1}{n-1}\right)\right]$$

Hence $E(S_n) = np - (n-r)(p-q) < np$ for $p > q$ and $V(S_n) = npq$.

8. Let n_1 letters 'A' and n_2 letters 'B' be arranged at random in a sequence. A run is a succession of like letters preceded and followed by none or an unlike letter. Let W be the total number of runs of 'A' s and 'B' s. Obtain expressions for Prob $\{W = r\}$, where r is a given positive even integer and also when r is odd.

Compute the expectation of W.

9. An urn contains K varieties of objects in equal numbers. The objects are drawn one at a time and replaced before the next drawing. Show that the probability that n and no less drawings will be required to produce objects of all varieties is

$$\sum_{r=0}^{k-1} (-1)^r \, {}^{k-1}C_r \left(\frac{k-1-r}{k}\right)^{n-1}$$

Hence or otherwise, find the expected number of drawings in a simple form.

10. An urn contains a white and b black balls. After a ball is drawn, it is to be returned to the urn if it is white, but if it is black, it is to be replaced by a white ball from another urn. Show that the probability of drawing a white ball after the foregoing operation has been repeated x times is

$$1-\frac{b}{a+b}\left(1-\frac{1}{a+b}\right)^x$$

11. A box contains k varieties of objects, the number of objects of each variety being the same. These objects are drawn one at a time and put back before the next drawing. Denoting by n the smallest number of drawings which produce objects of all varieties, find $E(n)$ and $V(n)$.

12. There is a lot of N objects from which objects are taken at random one by one with replacement. Prove that the expected value and variance of the least number of drawings needed to get n different objects are respectively given by

$$N\left[\frac{1}{N}+\frac{1}{N-1}+\ldots+\frac{1}{N-n+1}\right]$$

and $$N\left[\frac{1}{(N-1)^2}+\frac{2}{(N-2)^2}+\ldots+\frac{n-1}{(N-n+1)^2}\right]$$

13. A large population consits of equal number of individuals of c different types. Individuals are drawn at random one by one until at leasts one individual of each type has been found, whereupon sampling ceases. Show that the mean number of individuals in the sample is

$$c\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{c}\right)$$

and the variance of the number is

$$c^2\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\ldots+\frac{1}{c^2}\right)-c\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{c}\right)$$

14. (*a*) A point P is taken at random in a line AB of length $2a$, all positions of the point being equally likely. Show that the expected value of the area of the rectangle $AP.AB$ is $2a^2/3$ and the probability of the area exceeding $a^2/2$ $1/\sqrt{2}$.

(*b*) A point is chosen at random on a circle of radius a. Show that the expectation of its distance from another fixed point also on the circle is $4a/\pi$.

(*c*) Two points P and Q are selected at random in a square of side a. Prove that

$$E(|PQ|^2)=a^2/3$$

15. If the rools x_1, x_2 of the equation $x^2-ax+b=0$ are real and b is positive but otherwise unknown, prove that

$$E(x_i)=\tfrac{1}{6}a \text{ and } E(x_2)=\tfrac{5}{6}a$$

16. ***(Banach's Match-box Problem).*** A certain mathematician always carries two match boxes (initially containing N match-sides). Each time he wants a match-stick, he selects a box at random, inevitably a moment comes when he finds

a box empty. Show that the probability that there are exactly r match-sticks in one box when the other box becomes empty is

$$^{2N-r}C_N \times \frac{1}{2^{2N-r}}$$

Prove also that the expected number of matches is

$$^{2N}C_N \times \frac{2N+1}{2^{2N}} - 1$$

17. n couples procrete independetnly with no limits on family size. Births are single and independent and for the ith couple, the probability of a baby is p_i. The sex ratio S is defined as

$$S = \frac{\text{Mean number of all boys}}{\text{Mean number of all children}}$$

Show that if all couples,

(*i*) Stop procreating on the birth of a boy, then

$$S = n \Big/ \sum_{i=1}^{n} \frac{1}{p^i}$$

(*ii*) Stop procreating on birth of a girl, then

$$S = 1 - \left[n \Big/ \sum_{i=1}^{n} \frac{1}{q^i} \right], \text{ where } q_i = 1 - p_1$$

(*iii*) Stop procreating when they have children of both sexes, then

$$S = \left[\sum_{i=1}^{n} \frac{1}{q_i} - \sum_{i=1}^{n} p_i \right] \Big/ \left[\sum_{i=1}^{n} \frac{1}{p_i q_i} - n \right]$$

18. Show that if X is a random variable such that $P(a \le X \le b) = 1$, then $E(X)$ and Var (X) exist, and $a \le E(X) \le b$ and Var $(X) \le (b-a)^2/4$.

19. (X, Y) is a two-dimensional discrete random variable with the possible values 0 and 1 for X, and also 0 and 1 for Y, and with the joint probabilities given by

Y \ X	0	1
0	p_{00}	p_{10}
1	p_{01}	p_{11}

Find the characteristic functions $\psi_1(t)$, $\psi_2(t)$ and $\xi(t_1, t_2)$ for X, Y and (X, Y) respectively and show that $\psi(t_1, t_2) = \psi_1(t_1)\psi_2(t_2)$ when $p_{00}p_{11} = p_{01}p_{10}$.

20. For a given sequence $\{X_n\}$ of r.v.'s,

$$\varphi_n(t, X_n) = (\sin nt)/nt,$$

determine the distribution function of X_n. Hence show that even though the sequence of characteristic functions $\varphi_n(t)$ converges to a limit $\varphi(t)$, the sequence of distribution functions does not converge to a distribution function. What is the condition that is violated here? [Indian Civil Services, 1984]

21. Two continuous variates X and Y have a joint p.d.f. with a joint characteristic function $\varphi(t_1, t_2)$. If $g_X(x)$ is the marginal density, show that $\mu'_r(x)$, the rth simple moment for the conditional distribution of Y given $X = x$, satisfies the equation

$$\mu'_r(x) \cdot g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\partial^r \phi(t_1, 0)}{\partial t_2^r} e^{-it_1 x} dt_1$$

[Indian Civil Services, 1987]

22. Prove that the real part of a characteristic function is again a characteristic function. Prove further that if $\psi_1(t) = a_1(t) + ib_1(t)$ and $\psi_2(t) = a_2(t) + ib_2(t)$ are characteristic functions, then $a_1(t) a_2(t) - b_1(t) b_2(t)$ is a characteristic function.

23. Show that for any distribution

$$\int_{-\infty}^{\infty} \left(1 - \frac{x^2}{t^2}\right) dF(x) \le \int_{-t}^{t} dF(x)$$

and hence deduce $P[\,|X - E(X)| > k\sigma\,] \le \dfrac{1}{k^2}$, where $k > 0$ and $\text{Var}(X) = \sigma^2$.

24. (*a*) Let X be a random variable with moment generating function $M(t), -h < t < h$. Prove that

$$P(X \ge a) \le e^{-at} M(t),\ 0 < t < h$$

and that

$$P(X \le a) \le e^{-at} M(t), -h < t < 0.$$

(*b*) Let $f(x, y) = xe^{-x(y+1)}$; $x > 0, y > 0$
$= 0$, elsewhere

Find moment generating function of $Z = XY$.

Hint.
$$M_{XY}(t) = \int_0^{\infty}\int_0^{\infty} e^{txy} f(x, y)\, dx\, dy$$
$$= \int_0^{\infty} \left[x e^{-x} \left\{ \int_0^{\infty} e^{-(1-t)xy} dy \right\} \right] dx\ ;\ 1 - t > 0$$
$$= \int_0^{\infty} x e^{-x} \frac{1}{(1-t)x} dx = \frac{1}{1-t}\ ; t < 1.$$

25. The probability of obtaining a 6 with a biased die is p, where $(0 < p < 1)$. Three players A, B and C roll this die in order, A starting. The first one to throw a 6 wins. Find the probability of winning for A, B and C.

If X is a random variable which takes the value r if the game finishes at the rth throw, determine the probability generating function of X and hence, or otherwise, evluate $E(X)$ and Var (X).

Hint. Probabilities for the wins of A, B and C are $p/(1-q^3)$, $pq(1-q^3)$ and $pq^2/(1-q^3)$ respectively.

$$P(X=r)=pq^{r-1}, \text{ for } r \geq 1.$$

The probability generating function of X is $P(s)=ps/(1-qs)$,

whence $E(X)=1/p$ and Var $(X)=1/qp^2$.

26. Define convergence in probability. Let $X_1, X_2, \ldots$ be i.i.d. variates with $f(x) = e^{-(x-1)}$, $x \geq 1$. Show that $Y_n \rightarrow 1$ in probability where $Y_n = \text{Min}(X_k)$; $1 \leq k \geq n$. **(Indian Civil Services, 1982)**

27. Let $\{X_n, n=1,2,\ldots\}$ be a sequence of standardised variates and Corr $(X_m, X_n) = \exp[-|m-n|\alpha]$, $\alpha>0$ and $m \neq n$. Show that W.L.L.N. holds for this sequence. **(Indian Civil Services, 1988)**

28. From the probability generating function (p.g.f.) of two random variables X and Y given by

$$P(s,t)=\exp[-\lambda-\mu-b+\lambda s+\mu t+bst],$$

(*i*) obtain the marginal p.g.f.'s and identify them,

(*ii*) obtain the p.g.f. of $X+Y$ and $P(X+Y)=0$, and

(*iii*) interpret the case $b=0$.

CHAPTER SEVEN

Theoretical Discrete Probability Distributions

7·0. Introduction. In the previous chapters we have discussed in detail the frequency distributions. In the present chapter we will discuss theoretical discrete distributions in which variables are distributed according to some definite probability law which can be expressed mathematically. The present study will also enable us to fit a mathematical model or a function of the form $y = p(x)$ to the observed data.

We have already defined distribution function, mathematical expectation, m.g.f., characteristic function and moments. This prepares us for a study of theoretical distributions. This chapter is devoted to the study of univariate (except for the multinomial) distributions like Binomial, Poisson, Negative binomial Geometric, Hypergeometric, Multinomial and Power-series distributions.

7·1. Bernoulli Distribution. A random variable X which takes two values 0 and 1, with probabilities q and p respectively, *i.e.*, $P(X = 1) = p$, $P(X = 0) = q$, $q = 1 - p$ is called a *Bernoulli variate* and is said to have a Bernoulli distribution.

Remark. Sometimes, the two values are $+1, -1$ instead of 1 and 0.

7·1·1. Moments of Bernoulli distribution. The r^{th} moment about origin is

$$\mu_r' = E(X^r) = 0^r \cdot q + 1^r \cdot p = p \; ; r = 1, 2, \ldots \qquad \ldots(7\cdot 1)$$

$$\mu_1' = E(X) = p, \quad \mu_2' = E(X^2) = p$$

$$\mu_2 = \text{Var}(X) = p - p^2 = pq$$

The m.g.f. of Bernoulli variate is given by :

$$M_X(t) = e^{0 \cdot t} \times P(X = 0) + e^{1 \cdot t} \cdot P(X = 1) = q + pe^t \qquad \ldots(7\cdot 1a)$$

Remark. Degenerate Random Variable. Sometimes we may come across a variate X which is degenerate at a point 'c', say, so that : $P(X = c) = 1$ and $= 0$ otherwise, *i.e.*, the whole mass of the variable is concentrated at a single point 'c'.

$$\text{Since } P(X = c) = 1, \text{ Var}(X) = 0.$$

Thus a degenerate r.v. X is characterised by Var $(X) = 0$.

M.g.f. of degenerate r.v. is given by

$$M_X(t) = E(e^{tX}) = e^{tc}\, P(X = c) = e^{ct} \qquad \ldots(7\cdot 1b)$$

7·2. Binomial Distribution. Binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700 and was first published posthumously in 1713, eight years after his death). Let a random experiment be performed repeatedly and let the occurrence of an event in a trial be called a success and its non-occurrence a failure. Consider a set of n independent Bernoullian trials (n

being finite), in which the probability 'p' of success in any trial is constant for each trial. Then $q = 1 - p$, is the probability of failure in any trial.

The probability of x successes and consequently $(n - x)$ failures in n independent trials, in a specified order (say) $SSFSFFFS...FSF$ (where S represents success and F failure) is given by the compound probability theorem by the expression :

$$P(SSFSFFFS...FSF) = P(S)P(S)P(F)P(S)P(F)P(F)P(F)P(S) \times \ldots \times P(F)P(S)P(F)$$

$$= p \cdot p \cdot q \cdot p \cdot q \cdot q \cdot q \cdot p \ldots q \cdot p \cdot q$$

$$= \underbrace{p \cdot p \ldots p}_{x \text{ factors}} \quad \underbrace{q \cdot q \cdot q \ldots q}_{(n-x) \text{ factors}} = p^x q^{n-x}$$

But x successes in n trials can occur in $\binom{n}{x}$ ways and the probability for each of these ways is $p^x q^{n-x}$. Hence the probability of x successes in n trials in *any order* whatsoever is given by the addition theorem of probability by the expression:

$$\binom{n}{x} p^x q^{n-x}$$

The probability distribution of the number of successes, so obtained is called the *Binomial probability distribution*, for the obvious reason that the probabilities of 0, 1, 2, ..., n successes, *viz.*,

$q^n, \binom{n}{1} q^{n-1} p, \binom{n}{2} q^{n-2} p^2, \ldots, p^n$, are the successive terms of the binomial expansion $(q + p)^n$.

Definition. *A random variable X is said to follow binomial distribution if it assumes only non-negative values and its probability mass function is given by*

$$P(X = x) = p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x} \; ; x = 0, 1, 2, \ldots, n \; ; q = 1 - p \\ 0, \textit{ otherwise} \end{cases} \quad \ldots(7 \cdot 2)$$

The two independent constants n and p in the distribution are known as the *parameters* of the distribution. 'n' is also, sometimes, known as the degree of the binomial distribution.

Binomial distribution is a discrete distribution as X can take only the integral values, *viz.*, 0, 1, 2,..., n. Any variable which follows binomial distribution is known as *binomial variate*.

We shall use the notation $X \sim B(n, p)$ to denote that the random variable X follows binomial distribution with parameters n and p.

The probability $p(x)$ in (7·2) is also sometimes denoted by $b(x, n, p)$.

Remarks 1. This assignment of probabilities is permissible because

$$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (q + p)^n = 1$$

2. Let us suppose that n trials constitute an experiment. Then if this experiment is repeated N times, the *frequency function* of the binomial distribution is given by

$$f(x) = Np(x) = N\binom{n}{x} p^x q^{n-x}\,; x = 0, 1, 2, ..., n \qquad ...(7.3)$$

and the expected frequencies of 0, 1, 2, ..., n successes are the successive terms of the binomial expansion, $N(q+p)^n$, $q + p = 1$.

3. Binomial distribution is important not only because of its wide applicability, but because it gives rise to many other probability distributions. Tables for $p(x)$ are available for various values of n and p.

4. **Physical conditions for Binomial Distribution.** We get the binomial distribution under the following experimental conditions.

(i) Each trial results in two mutually disjoint outcomes, termed as success and failure.

(ii) The number of trials 'n' is finite.

(iii) The trials are independent of each other.

(iv) The probability of success 'p' is constant for each trial.

The problems relating to tossing of a coin or throwing of dice or drawing cards from a pack of cards with replacement lead to binomial probabability distribution.

Example 7·1. *Ten coins are thrown simultaneously. Find the probability of getting at least seven heads.*

Solution. p = Probability of getting a head $= \frac{1}{2}$

q = Probability of not getting a head $= \frac{1}{2}$

The probability of getting x heads in a random throw of 10 coins is

$$p(x) = \binom{10}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{10-x} = \binom{10}{x}\left(\frac{1}{2}\right)^{10}\,; x = 0, , 2, ..., 10$$

∴ Probability of getting at least seven heads is given by

$$P(X \geq 7) = p(7) + p(8) + p(9) + p(10)$$

$$= \left(\frac{1}{2}\right)^{10}\left\{\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right\}$$

$$= \frac{120 + 45 + 10 + 1}{1024} = \frac{176}{1024}.$$

Example 7·2. *A and B play a game in which their chances of winning are in the ratio 3 : 2. Find A's chance of winning at least three games out of the five games played.* **[Burdwan Univ. B.Sc. (Hons.), 1993]**

Solution. Let p be the probability that 'A' wins the game. Then we are given $p = 3/5 \Rightarrow q = 1 - p = 2/5$.

Hence, by binomial probability law, the probability that out of 5 games played, A wins 'r' games is given by :

$$P(X=r)=p(r)=\binom{5}{r}\cdot(3/5)^r\,(2/5)^{5-r}\,;\;r=0,1,2,\ldots,5$$

The required probability that 'A' wins at least three games is given by :

$$P(X\geq 3)=\sum_{r=3}^{5}\binom{5}{r}\frac{3^r\cdot 2^{5-r}}{5^5}$$

$$=\frac{3^3}{5^5}\left[\binom{5}{3}2^2+\binom{5}{4}.3\times 2+1.3^2\times 1\right]=\frac{27\times(40+30+9)}{3125}=0{\cdot}68$$

Example 7·3. *If m things are distributed among 'a' men and 'b' women, show that the probability that the number of things received by men is odd, is*

$$\frac{1}{2}\left[\frac{(b+a)^m-(b-a)^m}{(b+a)^m}\right]$$

(Nagpur Univ B.Sc., 1989, '93)

Solution. p = Probability that a thing is received by man $=\dfrac{a}{a+b}$, then

$q=1-p=1-\dfrac{a}{a+b}=\dfrac{b}{a+b}$, is the probability that a thing is received by woman.

The probability that out of m things exactly x are received by men and the rest by women, is given by

$$p(x)={}^mC_x\,p^x\,q^{m-x}\,;\;x=0,1,2,\ldots,m$$

The probability P that the number of things received by men is odd is given by

$$P=p(1)+p(3)+p(5)+\ldots={}^mC_1\cdot q^{m-1}\cdot p+{}^mC_3\cdot q^{m-3}\cdot p^3+{}^mC_5\cdot q^{m-5}\cdot p^5+\ldots$$

Now

$$(q+p)^m=q^m+{}^mC_1\cdot q^{m-1}\cdot p+{}^mC_2\cdot q^{m-2}p^2+{}^mC_3\cdot q^{m-3}\cdot p^3+{}^mC_4\cdot q^{m-4}\cdot p^4+\ldots$$

and

$$(q-p)^m=q^m-{}^mC_1\cdot q^{m-1}\cdot p+{}^mC_2\cdot q^{m-2}\cdot p^2-{}^mC_3\cdot q^{m-3}\cdot p^3+{}^mC_4\cdot q^{m-4}\cdot p^4-\ldots$$

$$\therefore\quad (q+p)^m-(q-p)^m=2\,[{}^mC_1\cdot q^{m-1}\cdot p+{}^mC_3\cdot q^{m-3}\cdot p^3+\ldots]=2P$$

But $\quad q+p=1$ and $q-p=\dfrac{b-a}{b+a}$

$$\therefore\quad 1-\left(\frac{b-a}{b+a}\right)^m=2P\;\Rightarrow\;P=\frac{1}{2}\left[\frac{(b+a)^m-(b-a)^m}{(b+a)^m}\right]$$

Example 7·4 *An irregular six faced die is thrown and the expectation that in 10 throws it will give five even numbers is twice the expectation that it will give four even numbers. How many times in 10,000 sets of 10 throws each, would you expect it to give no even number.* **(Gujarat Univ. B.Sc. 1988)**

Solution. Let p be the probability of getting an even number in a throw of a die. Then the probability of getting x even numbers in ten throws of a die is

$$P(X=x)=\binom{10}{x}p^x\,q^{10-x}\,;\;x=0,1,2\ldots10$$

We are given that

$$P(X = 5) = 2\,P(X = 4)$$

i.e.,
$$\binom{10}{5} p^5 q^5 = 2 \binom{10}{4} p^4 q^6$$

$$\Rightarrow \quad \frac{10\,!\,p}{5\,!\,5\,!} = 2\,\frac{10\,!\,q}{4\,!\,6\,!}$$

$$\Rightarrow \quad \frac{p}{5} = \frac{2q}{6} = \frac{q}{3}$$

$\therefore\ 3p = 5q = 5(1-p) \ \Rightarrow\ 8p = 5 \ \Rightarrow\ p = 5/8$ and $q = 3/8$

$$\therefore \quad P(X = x) = \binom{10}{x} \left(\frac{5}{8}\right)^x \left(\frac{3}{8}\right)^{10-x}$$

Hence the required number of times that in 10,000 sets of 10 throws each, we get no even number

$$= 10{,}000 \times P(X = 0) = 10{,}000 \times \left(\frac{3}{8}\right)^{10} = 1 \text{ (approx.)}$$

Example 7·5 *In a precision bombing attack there is a 50% chance that any one bomb will strike the target. Two direct hits are required to destroy the target completely. How many bombs must be dropped to give a 99% chance or better of completely destroying the target ?* **[Gauhati Univ. M.A., 1992]**

Solution. We have :

p = Probability that the bomb strikes the target = 50% = $\frac{1}{2}$. Let n be the number of bombs which should be dropped to ensure 99% chance or better of completely destroying the target. This implies that "probability that out of n bombs, at least two strike the target, is greater than 0·99".

Let X be a r.v. representing the number of bombs striking the target. Then $X \sim B(n, p = \frac{1}{2})$ with

$$p(x) = P(X = x) = \binom{n}{x}\left(\frac{1}{2}\right)^x \cdot \left(\frac{1}{2}\right)^{n-x} = \binom{n}{x}\left(\frac{1}{2}\right)^n ; x = 0, 1, ..., n$$

We should have :

$$P(X \geq 2) \geq 0{\cdot}99$$

$$\Rightarrow \quad [1 - P(X \leq 1)] \geq 0{\cdot}99$$

$$\Rightarrow \quad [1 - \{p(0) + p(1)\}] \geq 0{\cdot}99$$

$$\Rightarrow \quad 1 - \left\{\binom{n}{0} + \binom{n}{1}\right\}\left(\frac{1}{2}\right)^n \geq 0{\cdot}99$$

$$\Rightarrow \quad 0{\cdot}01 \geq \frac{1+n}{2^n} \quad \Rightarrow \quad 2^n \times (0{\cdot}01) \geq 1 + n$$

$$\Rightarrow \quad 2^n \geq 100 + 100\,n \qquad ...(*)$$

By trial method, we find that the inequality (*) is satisfied by $n = 11$. Hence the minimum number of bombs needed to destroy the target completely is 11.

Example 7·6. *A department in a works has 10 machines which may need adjustment from time to time during the day. Three of these machines are old, each having a probability of 1/11 of needing adjustment during the day, and 7 are new, having corresponding probabilities of 1/21.*

Assuming that no machine needs adjustment twice on the same day, determine the probabilits that on a particular day :

(i) just 2 old and no new machines need adjustment.

(ii) If just 2 machines need adjustment, they are of the same type.

(Nagpur Univ. B.E., 1989)

Solution. Let p_1 = Probability that an old machine needs adjustment $= 1/11$

$\therefore \quad q_1 = 1 - p_1 = 10/11$

and p_2 = Probability that a new machine needs adjustment = 1/21

$\therefore \quad q_2 = 1 - p_2 = 20/21$

Then $P_1(r)$ = Probability that 'r' old machines need adjustment

$$= {}^3C_r\, p_1^r\, q_1^{3-r} = {}^3C_r\,(10/11)^{3-r}\,(1/11)^r$$

and $P_2(r)$ = Probability that 'r' new machine need adjustment

$$= {}^7C_r\, p_2^r\, q_2^{7-r} = {}^7C_r\,(1/21)^r\,(20/21)^{7-r}$$

(*i*) The probability that just two old machines and no new machine need adjustment is given (by the compound probability theorem) by the expression :

$$P_1(2)\cdot P_2(0) = {}^3C_2(1/11)^2\cdot(10/11)\cdot(20/21)^7 = 0{\cdot}016$$

(*ii*) Similarly the probability that just 2 new machines and no old machine need adjustment is

$$P_1(0)\cdot P_2(2) = (10/11)^3\cdot{}^7C_2\,(1/21)^2\cdot(20/21)^5 = 0{\cdot}028$$

The probability that "If just two machines need adjustment, they are of the same type" is the same as the probability that "either just 2 old and no new or just 2 new and no old machines need adjustment".

$\therefore$ Required probability = 0·016 + 0·028 = 0·044

7·2·1 Moments. The first four moments about origin of binomial distribution are obtained as follows :

$$\mu_1' = E(X) = \sum_{x=0}^{n} x\binom{n}{x}p^x q^{n-x} = np\sum_{x=1}^{n}\binom{n-1}{x-1}p^{x-1}q^{n-x}\,*$$

$$= np(q+p)^{n-1} = np \qquad (\because q + p = 1)$$

Thus the mean of the binomial distribution is np.

$$*\quad \binom{n}{x} = \frac{n}{x}\cdot\binom{n-1}{x-1} = \frac{n}{x}\cdot\frac{n-1}{x-1}\cdot\binom{n-2}{x-2}$$
$$= \frac{n}{x}\cdot\frac{n-1}{x-1}\cdot\frac{n-2}{x-2}\binom{n-3}{x-3}, \text{ and so on.}$$

$$\mu_2' = E(X^2) = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \{x(x-1) + x\} \frac{n(n-1)}{x(x-1)} \cdot \binom{n-2}{x-2} p^x q^{n-x}$$

$$= n(n-1)p^2 \left[\sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} q^{n-x} \right] + np$$

$$= n(n-1)p^2 (q+p)^{n-2} + np = n(n-1)p^2 + np$$

$$\mu_3' = E(X^3) = \sum_{x=0}^{n} x^3 \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \{x(x-1)(x-2) + 3x(x-1) + x\} p^x q^{n-x}$$

$$= n(n-1)(n-2)p^3 \sum_{x=3}^{n} \binom{n-3}{x-3} p^{x-3} q^{n-x}$$

$$+ 3n(n-1)p^2 \sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} q^{n-x} + np$$

$$= n(n-1)(n-2)p^3 (q+p)^{n-3} + 3n(n-1)p^2 (q+p)^{n-2} + np$$

$$= n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$$

Similarly

$$x^4 = x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x$$

Let $x^4 = Ax(x-1)(x-2)(x-3) + Bx(x-1)(x-2) + Cx(x-1) + x$

By giving to x the values 1, 2 and 3 respectively, we find the values of arbitrary constants A, B and C. Therefore,

$$\mu_4' = E(X^4) = \sum_{x=0}^{n} x^4 \binom{n}{x} p^x p^{n-x}$$

$$= n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$$

[On simplification]

Central Moments of Binomial Distribution :

$$\mu_2 = \mu_2' - \mu_1'^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p) = npq$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= \{n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np\} - 3\{n(n-1)p^2 + np\}np + 2(np)^3$$

$$= np[-3np^2 + 3np + 2p^2 - 3p + 1 - 3npq]$$

$$= np[3np(1-p) + 2p^2 - 3p + 1 - 3npq]$$

$$= np\,[2p^2 - 3p + 1] = np\,(2p^2 - 2p + q) = npq\,(1 - 2p)$$
$$= npq\,[q + p - 2p] = npq\,(q - p)$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = npq\,[1 + 3\,(n - 2)\,pq]$$
[On simplification]

Hence

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{n^2p^2q^2(q-p)^2}{n^3p^3q^3} = \frac{(q-p)^2}{npq} = \frac{(1-2p)^2}{npq} \quad ...(7{\cdot}4)$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{npq\,\{1 + 3\,(n-2)\,pq\}}{n^2p^2q^2} = \frac{1 + 3\,(n-2)\,pq}{npq} = 3 + \frac{1-6pq}{npq} \quad ...(7{\cdot}5)$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}} = \frac{1-2p}{\sqrt{npq}},\ \gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq} \quad ...(7{\cdot}5\,a)$$

Example 7·7 *Comment on the following :*
The mean of a binomial distribution is 3 and variance is 4'.

Solution. If the given binomial distribution has parameters n and p, then we are given

$$\text{Mean} = np = 3 \quad ...(*)$$
and
$$\text{Variance} = npq = 4 \quad ...(**)$$

Dividing (**) by (*), we get $q = 4/3$,
which is impossible, since probability cannot exceed unity. Hence the given statement is wrong.

Example 7·8. *The mean and variance of binomial distribution are 4 and $\frac{4}{3}$ respectively. Find $P(X \geq 1)$.* (Sardar Patel Univ. B.Sc. 1993)

Solution. Let $X \sim B\,(n, p)$. Then we are given

$$\text{Mean} = E\,(X) = np = 4$$
and
$$\text{Var}\,(X) = npq = \tfrac{4}{3} \quad ...(*)$$

Dividing, we get

$$q = \tfrac{1}{3} \quad \Rightarrow \quad p = \tfrac{2}{3}$$

Substituting in (*), we get

$$n = \frac{4}{p} = \frac{4 \times 3}{2} = 6.$$

$$P(X \geq 1) = 1 - P\,(X = 0) = 1 - q^n = 1 - (1/3)^6 = 1 - (1/729)$$
$$= 1 - 0{\cdot}00137 = 0{\cdot}99863$$

Example 7·9 *If $X \sim B\,(n, p)$, show that :*

$$E\left(\frac{X}{n} - p\right)^2 = \frac{pq}{n};\ Cov\left(\frac{X}{n}, \frac{n-X}{n}\right) = -\frac{pq}{n}$$

(Delhi Univ. B.Sc., 1989)

Solution. Since $X \sim B(n, p)$, $E(X) = np$ and $\text{Var}(X) = npq$

$$\therefore \quad E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = p \,; \quad \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot \text{Var}(X) = \frac{pq}{n}$$

(i) $$E\left(\frac{X}{n} - p\right)^2 = E\left[\frac{X}{n} - E\left(\frac{X}{n}\right)\right]^2 = \text{Var}\left(\frac{X}{n}\right) = \frac{pq}{n}$$

(ii) $$\begin{aligned} \text{Cov}\left(\frac{X}{n}, \frac{n-X}{n}\right) &= E\left[\left\{\frac{X}{n} - E\left(\frac{X}{n}\right)\right\}\left\{\frac{n-X}{n} - E\left(\frac{n-X}{n}\right)\right\}\right] \\ &= E\left[\left(\frac{X}{n} - p\right)\left\{\left(1 - \frac{X}{n}\right) - (1-p)\right\}\right] \\ &= E\left[\left(\frac{X}{n} - p\right)\left\{-\left(\frac{X}{n} - p\right)\right\}\right] \\ &= -E\left(\frac{X}{n} - p\right)^2 = -\text{Var}\left(\frac{X}{n}\right) = -\frac{pq}{n} \end{aligned}$$

7·2·2 Recurrence Relation for the moments of Binomial Distribution.
(Renovsky Formula)

By def.,

$$\mu_r = E\{X - E(X)\}^r = \sum_{x=0}^{n} (x - np)^r \binom{n}{x} p^x q^{n-x}$$

Differentiating with respect to p, we get

$$\begin{aligned} \frac{d\mu_r}{dp} &= \sum_{x=0}^{n} \binom{n}{x} \Big[-nr(x - np)^{r-1} p^x q^{n-x} \\ &\qquad + (x - np)^r \{x p^{x-1} q^{n-x} - (n-x) p^x q^{n-x-1}\}\Big] \\ &= -nr \sum_{x=0}^{n} \binom{n}{x} (x - np)^{r-1} p^x q^{n-x} \\ &\qquad + \sum_{x=0}^{n} \binom{n}{x} (x - np)^r p^x q^{n-x} \left\{\frac{x}{p} - \frac{n-x}{q}\right\} \\ &= -nr \sum_{x=0}^{n} (x - np)^{r-1} p(x) + \sum_{x=0}^{n} (x - np)^r p(x) \frac{(x - np)}{pq} \\ &= -nr \sum_{x=0}^{n} (x - np)^{r-1} p(x) + \frac{1}{pq} \sum_{x=0}^{n} (x - np)^{r+1} p(x) \end{aligned}$$

$$\therefore \quad \frac{d\mu_r}{dp} = -nr\,\mu_{r-1} + \frac{1}{pq}\,\mu_{r+1}$$

$$\Rightarrow \quad \mu_{r+1} = pq\left[nr\,\mu_{r-1} + \frac{d\mu_r}{dp}\right] \qquad \ldots(7\cdot 6)$$

Putting $r = 1, 2$ and 3 successively in (7·6), we get

$$\mu_2 = pq\left[n\mu_o + \frac{d\mu_1}{dp}\right] = npq \qquad (\because \mu_0 = 1 \text{ and } \mu_1 = 0)$$

$$\mu_3 = pq\left[2n\mu_1 + \frac{d\mu_2}{dp}\right] = pq\cdot\frac{d(npq)}{dp} = npq\frac{d}{dp}\{p(1-p)\}$$

$$= npq\frac{d}{dp}(p-p^2) = npq(1-2p) = npq(q-p)$$

and
$$\mu_4 = pq\left[3n\mu_2 + \frac{d\mu_3}{dp}\right] = pq\left[3n\cdot npq + \frac{d}{dp}\{npq(q-p)\}\right]$$

$$= pq\left[3n^2pq + n\frac{d}{dp}\{p(1-p)(1-2p)\}\right]$$

$$= pq\left[3n^2pq + n\frac{d}{dp}(p-3p^2+2p^3)\right]$$

$$= pq[3n^2pq + n(1-6p+6p^2)] = pq[3n^2pq + n(1-6pq)]$$

$$= npq[3npq + 1 - 6pq] = npq[1 + 3pq(n-2)]$$

Example 7·10 *Show that the rth moment μ_r' about the origin of the binomial distribution of degree n is given by :*

$$\mu_r' = \left(p\frac{\partial}{\partial p}\right)^r (q+p)^n \qquad ...(*)$$ **[Patna Univ. B.Sc. (Hons.), 1993]**

Solution. We shall prove this result by using the principle of mathematical induction. We have

$$(q+p)^n = \sum_{x=0}^{n}\binom{n}{x}p^xq^{n-x} \Rightarrow \frac{\partial}{\partial p}(q+p)^n = \sum_{x=0}^{n}\binom{n}{x}q^{n-x}xp^{x-1}$$

$$\therefore \quad p\frac{\partial}{\partial p}(q+p)^n = p\sum_{x=0}^{n}\binom{n}{x}q^{n-x}xp^{x-1} = \sum_{x=0}^{n}\binom{n}{x}p^xq^{n-x}x = \mu_1'$$

Thus the result $(*)$ is true for $r = 1$.

Let us now assume that the result $(*)$ is true for $r = k$, so that

$$\left(p\frac{\partial}{\partial p}\right)^k (q+p)^n = \mu_k' = \sum_{x=0}^{n}\binom{n}{x}p^xq^{n-x}x^k \qquad ...(**)$$

Differentiate $(**)$ partially w.r. to p and multiply both sides by p to get :

$$p\left(\frac{\partial}{\partial p}\right)\left[\left(p\frac{\partial}{\partial p}\right)^k (q+p)^n\right] = \sum_{x=0}^{n}\binom{n}{x}p^xq^{n-x}x^{k+1} = E\left(X^{k+1}\right)$$

$$\Rightarrow \quad \left(p\frac{\partial}{\partial p}\right)^{k+1}(q+p)^n = \mu_{k+1}'$$

Hence if the result $(*)$ is true for $r = k$, it is also true for $r = k + 1$. It is already shown to be true for $k = 1$. Hence by the principle of mathematical induction, $(*)$ is true for all positive integral values of r.

7·2·3. Factorial Moments of Binomial Distribution. The rth factorial moment of the Binomial distribution is:

$$\mu_{(r)}' = E\{X^{(r)}\} = \sum_{x=0}^{n} x^{(r)} p(x) = \sum_{x=0}^{n} x^{(r)} \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= n^{(r)} p^r \sum_{x=r}^{n} \frac{(n-r)!}{(x-r)!(n-x)!} p^{x-r} q^{n-x} = n^{(r)} p^r (q+p)^{n-r}$$

$$= n^{(r)} p^r \qquad \ldots(7\cdot 7)$$

$$\mu_{(1)}' = E\{X^{(1)}\} = np = \text{Mean}$$

$$\mu_{(2)}' = E\{X^{(2)}\} = n^{(2)} p^2 = n(n-1)p^2$$

$$\mu_{(3)}' = E\{X^{(3)}\} = n^{(3)} p^3 = n(n-1)(n-2)p^3$$

Now $\mu_{(2)} = \mu_{(2)}' - \mu_{(1)}'^2 + \mu_{(1)}' = n^2p^2 - np^2 - n^2p^2 + np = npq$

$$\mu_{(3)} = \mu_{(3)}' - 3\mu_{(2)}' \mu_{(1)}' + 2\mu_{(1)}'^3 - 2\mu_{(1)}'$$

$$= n(n-1)(n-2)p^3 - 3n(n-1)p^2\, np + 2n^3p^3 - 2np = -2npq(1+p)$$

[On simplification]

7·2·4. Mean Deviation About Mean of Binomial Distribution. The mean deviation η about the mean np of the binomial distribution is given by

$$\eta = \sum_{x=0}^{n} |x - np|\, p(x) = \sum_{x=0}^{n} |x-np| \binom{n}{x} p^x q^{n-x},$$

(x being an integer)

$$= \sum_{x=0}^{np} -(x-np)\binom{n}{x} p^x q^{n-x} + \sum_{x=np}^{n} (x-np)\binom{n}{x} p^x q^{n-x}$$

$$= 2\sum_{x=np}^{n} (x-np)\binom{n}{x} p^x q^{n-x} \; *$$

$$= 2\sum_{\mu}^{n} (x-np)\binom{n}{x} p^x q^{n-x},$$

where μ is the greatest integer contained in $np + 1$.

$$= 2\sum_{\mu}^{n} \left[\{xq - (n-x)p\}\binom{n}{x} p^x q^{n-x}\right]$$

$$= 2\sum_{\mu}^{n} \left[\frac{n!}{(x-1)!(n-x)!} p^x q^{n-x+1} - \frac{n!}{x!(n-x-1)!} p^{x+1} q^{n-x}\right]$$

---

* $\because \quad \sum_{x=0}^{n} x\binom{n}{x} p^x q^{n-x} = np$

$$\Rightarrow \quad \sum_{x=0}^{n} (x-np)\binom{n}{x} p^x q^{n-x} = 0$$

$$= 2 \sum_{x=\mu}^{n} \left[t_{x-1} - t_x \right], \text{ where } t_x = \frac{n!}{x!\,(n-x-1)!} p^{x+1} q^{n-x}$$

$$= 2 \left[t_{\mu-1} - t_n \right] = 2\, t_{\mu-1}$$

This is obtained by summing over x and using $t_n = o$

$$\therefore \eta = 2t_{\mu-1} = 2 \frac{n!}{(\mu-1)!\,(n-\mu)!} \cdot p^{\mu} q^{n-\mu+1}$$

$$= 2npq \binom{n-1}{\mu-1} p^{\mu-1} q^{n-\mu} \qquad ...(7\cdot8)$$

7·2·5. Mode of the Binomial Distribution. We have

$$\frac{p(x)}{p(x-1)} = \binom{n}{x} p^x q^{n-x} \Big/ \binom{n}{x-1} p^{x-1} q^{n-x+1}$$

$$= \frac{n!}{(n-x)!\,x!} p^x q^{n-x} \Big/ \frac{n!}{(x-1)!\,(n-x+1)!} p^{x-1} q^{n-x+1}$$

$$= \frac{(n-x+1)p}{xq} = \frac{xq + (n-x+1)p - xq}{xq}$$

$$= 1 + \frac{(n+1)p - x(p+q)}{xq} = 1 + \frac{(n+1)p - x}{xq} \qquad ...(7\cdot9)$$

Mode is the value of x for which $p(x)$ is maximum.

We discuss the following two cases :

Case 1. When $(n+1)p$ is not an integer

Let $(n+1)p = m + f$, where m is an integer and f is fractional such that $0 < f < 1$. Substituting in (7·9), we get

$$\frac{p(x)}{p(x-1)} = 1 + \frac{(m+f) - x}{xq} \qquad ...(*)$$

From (*), it is obvious that

$$\frac{p(x)}{p(x-1)} > 1 \text{ for } x = 0, 1, 2, ..., m$$

and $$\frac{p(x)}{p(x-1)} < 1 \text{ for } x = m+1, m+2, ..., n$$

$$\Rightarrow \quad \frac{p(1)}{p(0)} > 1, \frac{p(2)}{p(1)} > 1, ..., \frac{p(m)}{p(m-1)} > 1,$$

and $$\frac{p(m+1)}{p(m)} < 1, \frac{p(m+2)}{p(m+1)} < 1, ..., \frac{p(n)}{p(n-1)} < 1,$$

$$\therefore p(0) < p(1) < p(2) < ... < p(m-1) < p(m) > p(m+1) > p(m+2) > p(m+3) ... > p(n),$$

Thus in this case there exists unique modal value for binomial distribution and it is m, the integral part of $(n+1)p$.

Case II. *When $(n+1)p$ is an integer.*

Let $(n+1)p = m$ *(an integer).*

Substituting in (7·9), we get

$$\frac{p(x)}{p(x-1)} = 1 + \frac{m-x}{xq} \qquad \ldots(^{**})$$

From (**) it is obvious that

$$\left.\frac{p(x)}{p(x-1)}\right\} \begin{array}{l} > 1 \text{ for } x = 1, 2, \ldots, m-1 \\ = 1 \text{ for } x = m \\ < 1 \text{ for } x = m+1, m+2, \ldots, n \end{array}$$

Now proceeding as in case 1, we have :

$p(0) < p(1) < \ldots < p(m-1) = p(m) > p(m+1) > p(m+2) > \ldots > p(n)$

Thus in this case the distribution is bimodal and the two modal values are m and $m-1$.

Example 7·11. *Determine the binomial distribution for which the mean is 4 and variance 3 and find its mode.* **(Madurai Kamraj Univ B.Sc. 1993)**

Solution, Let $X \sim B(n, p)$, then we are given that

$$E(X) = np = 4 \qquad ..(^*)$$

and

$$\text{Var}(X) = npq = 3 \qquad ..(^{**})$$

Dividing (**) by (*), we get

$$q = \tfrac{3}{4} \quad \Rightarrow \quad p = 1 - q = \tfrac{1}{4}$$

Hence from (*), $\quad n = \frac{4}{p} = 16$

Thus the given binomial distribution has parameters $n = 16$ and $p = 1/4$.

Mode. We have $(n+1)p = 4{\cdot}25$, which is not an integer. Hence the unique mode of the binomial distribution is 4, the integral part of $(n+1)p$.

Example 7·12. *Show that for $p = 0{\cdot}50$, the binomial distribution has a maximum probability at $X = \frac{1}{2}n$, if n is even, and at $X = \frac{1}{2}(n-1)$ as well as $X = \frac{1}{2}(n+1)$, if n is odd.* **(Mysore Univ., B. Sc. 1991)**

Solution. Here we have to find the mode of the binomial distribution.

(i) Let n be even $= 2m$, (say), $m = 1, 2, \ldots$

$\therefore$ If $p = 0{\cdot}5$, then $(n+1)p = (2m+1) \times \left(\frac{1}{2}\right) = m + 0{\cdot}5$

Hence in this case, the distribution is unimodal, the unique mode being at $X = m = n/2$.

(ii) Let n be odd $= (2m+1)$, say. Then

$(n+1)p = (2m+2) \times \frac{1}{2} = m + 1$ (Integer)

$$= \frac{n-1}{2} + 1 = \frac{n+1}{2}$$

Since $(n+1)p$ is an integer, the distribution is bimodal, the two modes being $\frac{1}{2}(n+1)$ and $\frac{1}{2}(n+1)-1 = \frac{1}{2}(n-1)$.

7·2·6. Moment Generating Function of Binomial Distribution. Let X be a variable following binomial distribution, then

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x q^{n-x} = \sum_{x=0}^{n} (pe^t)^x q^{n-x}\binom{n}{x} = (q+pe^t)^n \quad ...(7·10)$$

M.G.F. about Mean of Binomial Distribution :

$$E\left\{e^{t(X-np)}\right\} = E(e^{tx}e^{-tnp}) = e^{-tnp}\cdot E(e^{tX}) = e^{-tnp}\cdot M_X(t)$$

$$= e^{-tnp}(q+pe^t)^n = (qe^{-pt}+pe^{qt})^n \quad ...(7·11)$$

$$= \left[q\left\{1 - pt + \frac{p^2t^2}{2!} - \frac{p^3t^3}{3!} + \frac{p^4t^4}{4!} - \dots\right\}\right.$$

$$\left. + p\left\{1 + tq + \frac{t^2q^2}{2!} + \frac{t^3q^3}{3!} - \dots\right\}\right]^n$$

$$= \left[1 + \frac{t^2}{2!}pq + \frac{t^3}{3!}pq\,(q^2-p^2) + \frac{t^4}{4!}pq\,(q^3+p^3) + \dots\right]^n$$

$$= \left[1 + \left\{\frac{t^2}{2!}\cdot pq + \frac{t^3}{3!}\cdot pq\,(q-p) + \frac{t^4}{4!}pq\,(1-3pq) + \dots\right\}\right]^n$$

$$= \left[1 + \binom{n}{1}\left\{\frac{t^2}{2!}\cdot pq + \frac{t^3}{3!}pq\,(q-p) + \frac{t^4}{4!}pq\,(1-3pq) + \dots\right\}\right.$$

$$\left. + \binom{n}{2}\left\{\frac{t^2}{2!}pq + \frac{t^3}{3!}pq\,(q-p) + \dots\right\}^2 + \dots\right.$$

Now
$$\mu_2 = \text{Coefficient of } \frac{t^2}{2!} = npq$$

$$\mu_3 = \text{Coefficient of } \frac{t^3}{3!} = npq\,(q-p)$$

$$\mu_4 = \text{Coefficient of } \frac{t^4}{4!} = npq\,(1-3pq) + 3n(n-1)p^2q^2$$

$$= npq\,(1-3pq) + 3n^2p^2q^2 - 3np^2q^2$$

$$= 3n^2p^2q^2 + npq\,(1-6pq)$$

Example 7·13 *X is binomially distributed with parameters n and p. What is the distrbution of $Y = n - X$?* **[Delhi Univ. B.Sc. (Maths Hons.), 1990]**

Solution. $X \sim B(n,p)$, represents the number of successes in n independent trials with constant probability p of success for each trial.

$\therefore$ $Y = n - X$, represents the number of failures in n independent trial with constant probability 'q' of failure for each trial. Hence $Y = n - X \sim B(n, q)$

Aliter Since $X \sim B(n, p)$; $M_X(t) = E(e^{tX}) = (q + pe^t)^n$

$$\begin{aligned} \therefore \quad M_Y(t) &= E(e^{tY}) = E(e^{t(n-X)}) \\ &= e^{nt} \cdot E(e^{-tX}) = e^{nt} M_X(-t) \\ &= e^{nt} \cdot (q + pe^{-t})^n \\ &= \left[e^t \left(q + pe^{-t}\right)\right]^n = \left(p + qe^t\right)^n \end{aligned}$$

Hence by uniqueness theorem of m.g.f., $Y = n - X \sim B(n, q)$

Example 7·14. *The m.g.f. of a r.v. X is* $\left(\frac{2}{3} + \frac{1}{3}e^t\right)^9$. *Show that :*

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \sum_{x=1}^{5} \binom{9}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{9-x}$$

[Delhi Univ. B.Sc. (Maths Hons.), 1989]

Solution. Since $M_X(t) = \left(\frac{2}{3} + \frac{1}{3}e^t\right)^9 = \left(q + pe^t\right)^n$,

by uniqueness theorem of m.g.f. $X \sim B\left(n = 9, p = \frac{1}{3}\right)$

Hence $E(X) = \mu_x = np = 3$; $\sigma_X^2 = npq = 9 \times \frac{1}{3} \times \frac{2}{3} = 2$

$\mu \pm 2\sigma = 3 \pm 2 \times \sqrt{2} = 3 \pm 2 \times 1{\cdot}4 = (0{\cdot}2,\ 5{\cdot}8)$

$$\begin{aligned} \therefore \quad P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(0{\cdot}2 < X < 5{\cdot}8) = P(1 \leq X \leq 5) \\ &= \sum_{x=1}^{5} p(x) = \sum_{x=1}^{5} {}^nC_x p^x q^{n-x} \\ &= \sum_{x=1}^{5} {}^9C_x (1/3)^x (2/3)^{9-x} \end{aligned}$$

7·2·7. Additive Property of Binomial Distribution. Let $X \sim B(n_1, p_1)$ and $Y \sim B(n_2, p_2)$ be independent random variables. Then

$$M_X(t) = (q_1 + p_1 e^t)^{n_1}, \quad M_Y(t) = (q_2 + p_2 e^t)^{n_2} \qquad \ldots(*)$$

What is the distribution of $X + Y$?

We have

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \quad [\because X \text{ and } Y \text{ are independent}] \\ &= (q_1 + p_1 e^t)^{n_1} \cdot (q_2 + p_2 e^t)^{n_2} \qquad \ldots(**) \end{aligned}$$

Since (**) cannot be expressed in the form $(q + p e^t)^n$, from uniqueness theorem of m.g.f.'s it follows that $X + Y$ is not a binomial variate. Hence, *in general the sum of two independent binomial variates is not a binomial variate.*

In other words, binomial distribution does not possess the additive or reproductive property.

However, if we take $p_1 = p_2 = p$ (say), then from (**), we get

$$M_{X+Y}(t) = (q + pe^t)^{n_1+n_2},$$

which is the m.g.f. of a binomial variate with parameters $(n_1 + n_2, p)$. Hence, by uniqueness theorem of m.g.f.'s $X + Y \sim B(n_1 + n_2, p)$. Thus the *binomial distribution possesses the additive or reproductive property if* $p_1 = p_2$.

Generalisation. *If X_i, $(i = 1, 2, ..., k)$ are independent binomial variates with parameters (n_i, p), $(i = 1, 2, ..., k)$ then their sum* $\sum_{i=1}^{k} X_i \sim B\left(\sum_{i=1}^{k} n_i, p\right)$.

The proof is left as an exercise to the reader.

Example 7·15. *If the independent random variables X, Y are binomially distributed, respectively with $n = 3$, $p = 1/3$, and $n = 5$, $p = 1/3$, write down the probability that $X + Y \geq 1$.*

Solution. We are given

$$X \sim B(3, \tfrac{1}{3}) \text{ and } Y \sim B(5, \tfrac{1}{3}).$$

Since X and Y are independent binomial random variables, with $p_1 = p_2 = \frac{1}{3}$, by the additive property of binomial distribution, we get

$$X + Y \sim B(3 + 5, \tfrac{1}{3}), \text{ i.e., } X + Y \sim B(8, \tfrac{1}{3})$$

$$\therefore \quad P(X + Y = r) = {}^8C_r (\tfrac{1}{3})^r (\tfrac{2}{3})^{8-r} \qquad ...(*)$$

$$\begin{aligned}\text{Hence } P(X + Y \geq 1) &= 1 - P(X + Y < 1)\\ &= 1 - P(X + Y = 0)\\ &= 1 - (\tfrac{2}{3})^8\end{aligned}$$

7·2·8. Characteristic Function of Binomial Distribution.

$$\varphi_X(t) = E(e^{itX}) = \sum_{x=0}^{n} e^{itx} p(x) = \sum_{x=0}^{n} e^{itx} \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^{it})^x q^{n-x} = (q + pe^{it})^n \qquad ...(7·12)$$

7·2·9. Cummulants of the Binomial Distribution. Cumulant generating function is

$$K_X(t) = \log M_X(t) = \log (q + pe^t)^n = n \log (q + pe^t)$$

$$= n \log \left[q + p\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + ...\right)\right]$$

$$= n \log \left[1 + p\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + ...\right)\right]$$

$$= n\left[p\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \ldots\right) - \frac{p^2}{2}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^2\right.$$

$$\left. + \frac{p^3}{3}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^3 - \frac{p^4}{4}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^4 + \ldots\right]$$

Mean $= \kappa_1 =$ Coefficient of t in $K_X(t) = np$

$\mu_2 = \kappa_2 =$ Coefficient of $\frac{t^2}{2!}$ in $K_X(t) = n(p - p^2) = np(1-p) = npq$

The coefficient of t^3 in $K_X(t)$

$$= n\left[\frac{p}{3!} - \frac{p^2}{2!}\cdot 2 \cdot \frac{1}{2!} + \frac{p^3}{3}\right] = \frac{np}{3!}(1 - 3p + 2p^2)$$

$\therefore\ \kappa_3 =$ Coefficient of $\frac{t^3}{3!}$ in $K_X(t) = np\,(1 - 3p + 2p^2)$

$$= np\,(1-p)(1-2p) = npq(1 - p - p) = npq\,(q - p)$$

$\therefore \qquad \mu_3 = \kappa_3 = npq\,(q-p)$

The Coefficient of t^4 in $K_X(t)$

$$= n\left[\frac{p}{4!} - \frac{p^2}{2!}\left(\frac{2}{3!} + \frac{1}{4}\right) + \frac{p^3}{3}\cdot\frac{3}{2!} - \frac{p^4}{4}\right]$$

$$= \frac{np}{4!}[1 - 7p + 12p^2 - 6p^3]$$

$\therefore\ \kappa_4 =$ Co efficient of $\frac{t^4}{4!}$ in $K_X(t) = np\,(1-p)(1 - 6p + 6p^2)$

$$= npq\,[1 - 6p\,(1-p)] = npq\,(1 - 6pq)$$

$\therefore\ \mu_4 = \kappa_4 + 3\kappa_2^2 = npq\,(1 - 6pq) + 3n^2p^2q^2$

$$= npq\,(1 - 6pq + 3npq) = npq\,[1 + 3pq\,(n-2)]$$

7·2·10. Recurrence Relation for Cumulants of Binomial Distribution. By def.,

$$\kappa_r = \left[\frac{d^r}{dt^r}\log M_X(t)\right]_{t=0} = n\left[\frac{d^r}{dt^r}\log(q + pe^t)\right]$$

$$\frac{d\kappa_r}{dp} = n\left[\frac{d^r}{dt^r}\cdot\frac{d}{dp}\log(q + pe^t)\right]_{t=0} = n\left[\frac{d^r}{dt^r}\cdot\frac{(-1 + e^t)}{q + pe^t}\right]_{t=0}$$

$$\kappa_{r+1} = n\left[\frac{d^{r+1}}{dt^{r+1}}\log(q + pe^t)\right]_{t=0}$$

$$= n\left[\frac{d^r}{dt^r}\cdot\frac{d}{dt}\log(q + pe^t)\right]_{t=0} = n\left[\frac{d^r}{dt^r}\left(\frac{pe^t}{q + pe^t}\right)\right]_{t=0}$$

$$= n\left[\frac{d^r}{dt^r}\left(1-\frac{q}{q+pe^t}\right)\right]_{t=0} = -nq\left[\frac{d^r}{dt^r}\left(\frac{1}{q+pe^t}\right)\right]_{t=0}$$

Hence

$$\kappa_{r+1} - pq\frac{d\kappa_r}{dp} = -nq\cdot\left[\frac{d^r}{dt^r}\left(\frac{1}{q+pe^t}\right)\right]_{t=0} - npq\left[\frac{d^r}{dt^r}\left(\frac{e^t-1}{q+pe^t}\right)\right]_{t=0}$$

$$= -nq\left[\frac{d^r}{dt^r}\left\{\frac{1+pe^t-p}{q+pe^t}\right\}\right]_{t=0}$$

$$= -nq\left[\frac{d^r}{dt^r}\left\{\frac{q+pe^t}{q+pe^t}\right\}\right]_{t=0} = -nq\left[\frac{d^r}{dt^r}(1)\right]_{t=0} = 0$$

$$\therefore \qquad \kappa_{r+1} = pq\frac{d\kappa_r}{dp} \qquad \ldots(7\cdot13)$$

In particular,

$$\kappa_2 = pq\cdot\frac{d\kappa_1}{dp} = pq\cdot\frac{d}{dp}(np) = npq \qquad (\because \kappa_1 = \text{mean} = np)$$

$$\kappa_3 = pq\cdot\frac{d\kappa_2}{dp} = pq\cdot\frac{d(npq)}{dp} = npq(q-p)$$

$$\kappa_4 = pq\cdot\frac{d\kappa_3}{dp} = pq\cdot\frac{d}{dp}\{npq(q-p)\}$$

$$= npq\frac{d}{dp}\{p(1-p)(1-2p)\}$$

$$= npq\cdot\frac{d}{dp}(p-3p^2+2p^3) = npq(1-6p+6p^2)$$

$$= npq[1-6p(1-p)] = npq(1-6pq)$$

7·2·11. Probability Generating Function of Binomial Distribution

$$P(s) = \sum_{\kappa=0}^{n} P(X=k)\, s^\kappa = \sum_{\kappa=0}^{n}\binom{n}{k}(ps)^\kappa q^{n-k} = (ps+q)^n \qquad \ldots(7\cdot13\,a)$$

The fact that this generating function is nth power of $(q + ps)$ shows that $p(x) = \{b(x; n, p)\}$ is the distribution of the sum $S_n = X_1 + X_2 + \ldots + X_n$ of n random variables with the common generating function $(q + ps)$. Each variable X_i assumes the value 0 with probability q and 1 with probability p.

Thus $$\{b(k; n, p)\} = \{b(k; 1, p)\}^{n*} \qquad \ldots(7\cdot13b)$$

Let X and Y be two independent random variables having $b(k; m, p)$ and $b(k; n, p)$ as their distributions, then

$$P_X(s) = (q+ps)^m \text{ and } P_Y(s) = (q+ps)^n$$

$$\therefore \qquad P_{X+Y}(s) = (q+ps)^m(q+ps)^n = (q+ps)^{m+n}$$

$$\therefore \{b(k; m, p)\} * \{b(k; n, p)\} = \{b(k; m+n, p)\} \qquad \ldots(7\cdot13c)$$

Also $\mu_{(1)}' = [\,n\,(q + ps)^{n-1}\,p\,]_{s=1} = np$

$\mu_{(2)}' = [\,n\,(n-1)\,(q+ps)^{n-2}\,p^2\,]_{s=1} = n\,(n-1)\,p^2$ and so on.

$\mu_{(r)}' = [n\,(n-1)\ldots(n-r+1)\,(q+ps)^{n-r}\,p^r\,]_{s=1}$

$= n\,(n-1)\ldots(n-r+1)\,p^r$

Example 7·16 *Show that*

$$E\left(\frac{1}{X+a}\right) = \int_0^1 t^{a-1}\,G(t)\,dt, \quad a > 0 \qquad \ldots(*)$$

where $G(t)$ is the probability generating function of X.

Find it when $X \sim B(n,p)$, and $a = 1$

[Delhi Univ. (Stat Hons.) Spl Course, 1988]

Solution. R.H.S. $= \int_0^1 t^{a-1}\cdot G(t)\,dt = \int_0^1 t^{a-1}\left(Et^X\right)dt$

$$= \int_0^1 \left\{t^{a-1}\left(\sum_x p_x t^x\right)\right\} dt = \sum_x \left[p_x \int_0^1 t^{x+a-1}\,dt\right]$$

$$= \sum_x p_x \cdot \frac{1}{(x+a)} = E\left(\frac{1}{X+a}\right) \qquad \ldots(**)$$

If $X\ B(n,p)$, then $G(t) = \sum_{x=0}^{n} t^x p_x = (q+pt)^n$

Hence taking $a = 1$ in (*) and using (**), we get :

$$E\left[\frac{1}{(X+a)}\right] = \int_0^1 (q+pt)^n\,dt = \left|\frac{(q+pt)^{n+1}}{(n+1)\,p}\right|_0^1 = \frac{1-q^{n+1}}{(n+1)\,p}$$

7·2·12. Recurrence Relation for the Probabilities of Binomial Distribution. *(Fitting of Binomial Distribution).*

We have

$$\frac{p\,(x+1)}{p\,(x)} = \frac{\binom{n}{x+1} p^{x+1}\,q^{n-x-1}}{\binom{n}{x} p^x\,q^{n-x}}$$

$$= \frac{n-x}{x+1}\cdot\frac{p}{q} \qquad \text{(On simplification)}$$

$$p\,(x+1) = \left\{\frac{n-x}{x+1}\cdot\frac{p}{q}\right\}p\,(x), \qquad \ldots(7\cdot14)$$

which is the required recurrence formula.

This formula provides us a very convenient method of graduating the given data by a binomial distribution. The only probability we need to calculate is $p\,(0)$

which is given by $p(0) = q^n$, where q is estimated from the given data by equating the mean $\bar{x}$ of the distribution to np, the mean of the binomial distribution. Thus $\hat{p} = \bar{x}/n$.

The remaining probabilities, *viz.*, $p(1), p(2), \ldots$ can now be easily obtained from (7·14) as explained below :

$$p(1) = [p(x+1)]_{x=0} = \left(\frac{n-x}{x+1} \cdot \frac{p}{q}\right)_{x=0} p(0)$$

$$p(2) = [p(x+1)]_{x=1} = \left(\frac{n-x}{x+1} \cdot \frac{p}{q}\right)_{x=1} p(1)$$

$$p(3) = [p(x+1)]_{x=2} = \left(\frac{n-x}{x+1} \cdot \frac{p}{q}\right)_{x=2} p(2)$$

and so on.

Example 7·17. *Seven coins are tossed and number of heads noted. The experiment is repeated 128 times and the following distribution is obtained:*

No. of heads	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Total*
Frequencies	*7*	*6*	*19*	*35*	*30*	*23*	*7*	*1*	*128*

Fit a Binomial distribution assuming

(i) The coin is unbaised,

(ii) The nature of the coin is not known.

(iii) Probability of a head for four coins is 0·5 and for the remaining three coins is 0·45.

Solution. In fitting Binomial distribution, first of all the mean and variance of the data are equated to np and npq respectively. Then the expected frequencies are calculated from these values of n and p. Here $n = 7$ and $N = 128$.

Case I. When the coin is unbaised

$$p = q = \tfrac{1}{2}, \ (p/q = 1)$$

Now
$$p(0) = q^n = (\tfrac{1}{2})^7 = (1/128)$$

$$f(0) = Nq^n = 128\,(\tfrac{1}{2})^7 = 1$$

Using the recurrence formula, the various probabilities, *viz.*, $p(1), p(2), \ldots$ can be easily calculated as shown below.

x	$\dfrac{n-x}{x+1}$	$\dfrac{n-x}{x+1} \cdot \dfrac{p}{q}$	*Expected frequency* $f(x) = Np(x)$
0	7	7	$f(0) = Np(0) = 1$
1	3	3	$f(1) = 1 \times 7 = 7$

2	$\frac{5}{3}$	$\frac{5}{3}$	$f(2) = 7 \times 3 = 21$
3	1	1	$f(3) = 21 \times \frac{5}{3} = 35$
4	$\frac{3}{5}$	$\frac{3}{5}$	$f(4) = 35 \times 1 = 35$
5	$\frac{1}{3}$	$\frac{1}{3}$	$f(5) = 35 \times \frac{3}{5} = 21$
6	$\frac{1}{7}$	$\frac{1}{7}$	$f(6) = 21 \times \frac{1}{3} = 7$
7			$f(7) = 7 \times \frac{1}{7} = 1$

Case II. When the nature of the coin is not known, then

$$np = \frac{1}{N} \sum_{i=1}^{n} f_i x_i = \frac{433}{128} = 3\cdot3828\,;\ n = 7$$

$\therefore$ $p = 0\cdot48326$ and $q = 0\cdot51674$, $(p/q = 0\cdot93521)$

$f(0) = Nq^7 = 128\,(0\cdot5167)^7 = 1\cdot2593$ (using logarithms)

x	$\frac{n-x}{x+1}$	$\frac{n-x}{x+1}\cdot\frac{p}{q}$	Expected frequency $f(x) = Np(x)$
0	7	6·54647	$f(0) = Np(0) = 1\cdot2593 \simeq 1$
1	3	2·80563	$f(1) = 1\cdot2593 \times 6\cdot54647 = 8\cdot2438 \simeq 8$
2	$\frac{5}{3}$	1·55868	$f(2) = 2\cdot80563 \times 8\cdot2438 = 23\cdot129 \simeq 23$
3	1	0·93521	$f(3) = 1\cdot55868 \times 23\cdot129 = 36\cdot05 \simeq 36$
4	$\frac{3}{5}$	0·56113	$f(4) = 0\cdot93521 \times 36\cdot05 = 33\cdot715 \simeq 34$
5	$\frac{1}{3}$	0·31174	$f(5) = 0\cdot56113 \times 33\cdot715 = 18\cdot918 \simeq 19$
6	$\frac{1}{7}$	0·13360	$f(6) = 0\cdot31174 \times 18\cdot918 = 5\cdot897 \simeq 6$
7			$f(7) = 0\cdot13360 \times 5\cdot897 = 0\cdot788 \simeq 1$

The probability generating functions (p.g.f.), say $P_X(s)$ for the 4 coins and $P_Y(S)$ for the remaining 3 coins are given by

$$P_X(s) = (0\cdot50 + 0\cdot50\,s)^4,\ P_Y(s) = (0\cdot55 + 0\cdot45s)^3 \quad \text{...[c.f. 7·13 (a)]}$$

Since all the throws are independent, the p.g.f. $P_{X+Y}(s)$ for the whole experiment is given by

$$P_{X+Y}(s) = P_X(s)\,P_Y(s) \qquad \text{...[}c.f.\ 7\cdot 13\ (b)\text{]}$$
$$= (0\cdot 50 + 0\cdot 50\, s)^4\,(0\cdot 55 + 0\cdot 45\, s)^3$$
$$= (0\cdot 0625 + 0\cdot 25\, s + 0\cdot 375\, s^2 + 0\cdot 25\, s^3 + 0\cdot 0625\, s^4)$$
$$\times\ (0\cdot 166375 + 0\cdot 408375\, s + 0\cdot 334125\ s^2 + 0\cdot 091125\, s^3)$$

Now $f(x) = N \times$ coefficient of t^x in $P_{X+Y}(t)$

$\therefore\quad f(0) = 128 \times \cdot 0625 \times \cdot 16637 = 1\cdot 13310$

$f(1) = 128\,\{\cdot 25 + \cdot 166375 + \cdot 408375 \times \cdot 0625\} = 8\cdot 5910$

$f(2) = 128\,\{\cdot 28396\} = 36\cdot 3470 \qquad f(5) = 128\,\{\cdot 14602\} = 18\cdot 6934$

$f(3) = 128\,\{\cdot 184117\} = 23\cdot 5669 \qquad f(6) = 128\,\{\cdot 04366\} = 5\cdot 5889$

$f(4) = 128\,\{\cdot 260570\} = 33\cdot 3529 \qquad f(7) = 128\,\{\cdot 005695\} = \cdot 72896$

Example 7·18. *Let X and Y be independent binomial variates, each with parameters n and p. Find P (X − Y = k).* **(Calcutta Univ. B.Sc., 1993)**

Solution. Since each of the variables X and Y takes the values $0, 1, 2, \ldots, n$, $Z = X - Y$ takes on the values $-n, -(n-1), \ldots, -1, 0, 1, \ldots, n$

$$P(Z = k) = \sum_{r=0}^{n} P(X = k + r \cap Y = r)$$

$$= \sum_{r=0}^{n} P(X = k + r)\cdot P(Y = r) \quad (\because X \text{ and } Y \text{ are independent}).$$

$$= \sum_{r=0}^{n} \binom{n}{k+r} p^{k+r}\cdot q^{n-k-r} \binom{n}{r} p^r q^{n-r}$$

$$= \sum_{r=0}^{n} \binom{n}{k+r}\binom{n}{r} p^{2r+k}\ q^{2n-2r-k} \qquad \ldots(*)$$

where $k = -n, -(n-1), \ldots, -2, -1, 0, 1, 2, \ldots, n$; and $q = 1 - p$.

In particular, we have :

$$P(Z = 0) = \sum_{r=0}^{n} \binom{n}{r}^2 \cdot p^{2r}\ q^{2n-2r}$$

$$P(Z = -n) = \sum_{r=0}^{n} \binom{n}{-n+r}\binom{n}{r} p^{2r-n}\ q^{3n-2r} = p^n q^n,$$

because we get the result when $r = n$ and for other values of $r < n$, $\binom{n}{-n+r}$ is not defined and hence taken as 0.

Example 7·19. *Find the m.g.f. of standard binomial variate $(X - np)/\sqrt{npq}$ and obtain its limiting form as $n \to \infty$. Also interpret the result.*

[Delhi Univ. B.Sc. (Stat. Hons.) 1990, 85]

Solution. We know that if $X \sim B(n, p)$, then

$$M_X(t) = (q + p\,e^t)^n$$

The m.g.f. of standard binomial variate.

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{X - \mu}{\sigma}, \text{(say)}$$

where $\mu = np$ and $\sigma^2 = npq$, is given by

$$M_Z(t) = e^{-\mu t/\sigma} M_X(t/\sigma)$$

$$= e^{-npt/\sqrt{npq}} \cdot (q + p\,e^{t/\sqrt{npq}})^n \qquad \text{[From (**)]}$$

$$= \left[e^{-pt/\sqrt{npq}} (q + p\,e^{t/\sqrt{npq}}) \right]^n$$

$$= \left[qe^{-pt/\sqrt{npq}} + pe^{qt/\sqrt{npq}} \right]^n$$

$$= \left[q \left\{ 1 - \frac{pt}{\sqrt{npq}} + \frac{p^2 t^2}{2npq} + 0'(n^{-3/2}) \right\} + p \left\{ 1 + \frac{qt}{\sqrt{npq}} + \frac{q^2 t^2}{2npq} + 0''(n^{-3/2}) \right\} \right]^n$$

where $0'(n^{-3/2})$ and $0''(n^{-3/2})$ involve terms containing $n^{3/2}$ and higher powers of n in the denominator.

$$\therefore\ M_Z(t) = \left[(q + p) + \frac{t^2 pq}{2npq}(p + q) + 0(n^{-3/2}) \right]^n$$

$$= \left[1 + \frac{t^2}{2n} + 0(n^{-3/2}) \right]^n$$

where $0(n^{-3/2})$ involves terms with $n^{3/2}$ and higher powers of n in the denominator.

$$\therefore\ \log M_Z(t) = n \log \left[1 + \frac{t^2}{2n} + 0(n^{-3/2}) \right]$$

$$= n \left[\left\{ \frac{t^2}{2n} + 0(n^{-3/2}) \right\} - \frac{1}{2} \left\{ \frac{t^2}{2n} + 0(n^{-3/2}) \right\}^2 + \ldots\ldots \right]$$

$$= \frac{t^2}{2} + 0'''(n^{-1/2})$$

where $0'''(n^{-1/2})$ involve terms with $n^{1/2}$ and higher powers of n in the denominator. Proceeding to the limit as $n \to \infty$, we get

$$\lim_{n \to \infty} \log M_Z(t) = \frac{t^2}{2}$$

$$\Rightarrow \qquad \lim_{n \to \infty} M_Z(t) = \exp(t^2/2) \qquad \ldots(**)$$

Interpretation. (**) is the m.g.f. of standard normal variate [c.f. **Remark** to § 8·2·5]. Hence by uniqueness theorem of moment generating functions, standard binomial variate tends to standard normal variate as $n \to \infty$. In other words, binomial distribution tends to normal distribution as $n \to \infty$.

Example 7·20. *A drunk performs a random walk over positions 0, ± 1, ± 2, ..., as follows. He starts at 0. He takes successive one unit steps, going to the right with probability p and to the left with probability (1 − p). His steps are independent. Let X denote his position after n steps. Find the distribution of (X + n)/2 and find E (X).* **(I.I.T. B.Tech., Dec. 1991)**

Solution. With the ith step of the drunk, let us associate a variable X_i defined as follows :

$X_i = 1$, if he takes the step to the right

$= -1$ if he takes the step to the left

Then $X = X_1 + X_2 + \ldots + X_n$, gives the position of the drunkard after n steps.

Define $\quad Y_i = (X_i + 1)/2$

Then $\quad Y_i = (1 + 1)/2 = 1$, with probability p

$\quad = (-1 + 1)/2 = 0$, with probability $1 - p = q$, (say).

Since the n steps of drunkard are independent, Y_i's, $(i = 1, 2, \ldots n)$ are i.i.d. Bernonlli variates with parameter p.

Hence $\sum\limits_{i=1}^{n} Y_i \sim B(n, p)$

$$\Rightarrow \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \left(\frac{X_i + 1}{2}\right) = \frac{1}{2}\left[\sum_{i=1}^{n} X_i + n\right] = \frac{X + n}{2} \sim B(n, p)$$

where $X = \sum\limits_{i=1}^{n} X_i$, is the position of the drunkard after n steps.

Since $(X + n)/2 \sim B(n, p)$, we have

$$E\left[\frac{X + n}{2}\right] = np \Rightarrow \frac{1}{2} E(X + n) = np$$

$$\Rightarrow E(X) + n = 2np \Rightarrow E(X) = n(2p - 1)$$

Example 7·21. *Suppose that the r.v. X is uniformly distributed on (0,1)*

i.e., $f_X(x) = 1 ; 0 \le x \le 1.$ ...(*)

Assume that the conditional distributional $Y | X = x$ *has a binomial distribution with parameters n and p = x, i.e.,*

$$P(Y = y | X = x) = \binom{n}{y} x^y (1 - x)^{n-y} ; y = 0, 1, 2, \ldots, n \qquad (**)$$

Find (a) E (Y)

(b) Find the distribution of Y. **(Punjab P.C.S., 1990)**

Solution. (a) We are given that the conditional distribution of

$Y | X = x \sim B(n, x)$ ...(i)

$\therefore E(Y | X = x) = nx$ ...(ii)

We have :

$$E(Y) = E[E(Y|X)] = E[nX] = nE(X) \qquad \text{[On using (ii)]}$$

$$\text{Now } E(X) = \int_0^1 x f(x)\,dx = \int_0^1 x\,dx = \frac{1}{2}.$$

$$\therefore\ E(Y) = n \times (\tfrac{1}{2}) = \frac{1}{2} n$$

(b) We have : $f_{X,Y}(x, y) = f_X(x) \cdot f_{Y|X}(y\,|\,x)$

Since X has (continuous) uniform distribution on (0,1) marginal distribution of Y is given by.

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx = \int_0^1 f_{Y|X}(y\,|\,x) \cdot f_X(x)\,dx$$

$$= \int_0^1 {}^nC_y \cdot x^y (1-x)^{n-y} \cdot 1 \cdot dx \qquad \text{[using (*) and (**)]}$$

$$= {}^nC_y \int_0^1 x^y (1-x)^{n-y}\,dx$$

$$= {}^nC_y \cdot \beta(y+1, n-y+1) = \frac{n!}{y!\,(n-y)!}\ \frac{\Gamma(y+1)\,\Gamma(n-y+1)}{\Gamma(n+2)}$$

$$= \frac{n!}{y!\,(n-y)!} \times \frac{y!\,(n-y)!}{(n+1)!}$$

$$= \frac{1}{n+1}\ ; \qquad y = 0, 1, 2, ..., n$$

Since Y takes the values 0, 1, 2, ..., n each with equal probability $1/(n+1)$, Y has discrete uniform distribution.

Remark We could find $E(Y)$ on using the distribution of Y in (b).

$$E(Y) = \sum_{y=0}^{n} y\,p(y) = \frac{1}{n+1} \sum_{y=0}^{n} y$$

$$= \frac{1}{n+1}[0 + 1 + 2 + ... + n] = \frac{n}{2},$$

as in Part (a).

Example 7·22. *If K (t) is the cumulative function about the origin of the Binomial Distribution of size n, show that*

$$\frac{d}{dt} K(t) = n\left\{1 + e^{-(z+t)}\right\}^{-1}, \text{ where } z = \log_e (p/q)$$

(b) *By expanding the R.H.S. in powers of t by Taylor's Theorem, show that*

$$\kappa_r = n\,\frac{d^{r-1}p}{dz^{r-1}}, \text{ where } \kappa_r \text{ is the } r\text{th cumulant.}$$

(c) *Hence or otherwise obtain the recurrence relation*

$$\kappa_{r+1} = pq \cdot \frac{d\kappa_r}{dp}, \quad r > 1$$

[Baroda Univ. B.Sc. 1993; Delhi Univ. B.Sc. (Stat. Hons.) 1992]

(d) *Prove that* $\kappa_{r+1} = \dfrac{d\kappa_r}{dz}$, *where* $z = \log_e (p/q)$

Solution. For binomial distribution with parameters n and p, we have

$$K(t) = \log M(t) = n \log (q + pe^t)$$

(a) $$\frac{d}{dt} K(t) = \frac{npe^t}{q + pe^t} = n\left(1 + \frac{q}{p} e^{-t}\right)^{-1}$$

if $z = \log_e (p/q) \Rightarrow (p/q) = e^z \Rightarrow (q/p) = e^{-z}$, then

$$\frac{d}{dt} K(t) = n[1 + e^{-(z+t)}]^{-1} \qquad \ldots(*)$$

(b) $$\kappa_r = \left[\frac{d^r}{dt^r} \kappa(t)\right]_{t=0} = \left[\frac{d^{r-1}}{dt^{r-1}} \cdot \frac{d}{dt} \kappa(t)\right]_{t=0}$$

$$= n\left[\frac{d^{r-1}}{dt^{r-1}} \left\{1 + e^{-(z+t)}\right\}^{-1}\right]_{t=0} = n\left[\frac{d^{r-1}}{dt^{r-1}} \left(\frac{e^{z+t}}{1 + e^{z+t}}\right)\right]_{t=0} \qquad \ldots(**)$$

By summetry of the function $e^{z+t}/(1 + e^{z+t})$ in t and z we have

$$\frac{d}{dt}\left(\frac{e^{z+t}}{1 + e^{z+t}}\right) = \frac{d}{dz}\left(\frac{e^{z+t}}{1 + e^{z+t}}\right)$$

$$\Rightarrow \quad \frac{d^{r-1}}{dt^{r-1}}\left(\frac{e^{z+t}}{1 + e^{z+t}}\right) = \frac{d^{r-1}}{dz^{r-1}}\left(\frac{e^{z+t}}{1 + e^{z+t}}\right)$$

Substituting in (**), we get

$$\kappa_r = n\left[\frac{d^{r-1}}{dz^{r-1}}\left(\frac{e^{z+t}}{1 + e^{z+t}}\right)\right]_{t=0} = n\frac{d^{r-1}}{dz^{r-1}}\left(\frac{e^z}{1 + e^z}\right)$$

$$= n\frac{d^{r-1}}{dz^{r-1}}(1 + e^{-z})^{-1} = n\frac{d^{r-1}}{dz^{r-1}}\left(1 + \frac{q}{p}\right)^{-1}$$

$$= n\frac{d^{r-1}p}{dz^{r-1}} \qquad \ldots(***)$$

(c) $$\frac{d\kappa_r}{dp} = n\frac{d}{dp}\left(\frac{d^{r-1}p}{dz^{r-1}}\right) = n\frac{d}{dz}\left(\frac{d^{r-1}p}{dz^{r-1}}\right)\frac{dz}{dp}$$

$$= n\frac{d^r p}{dz^r} \cdot \frac{1}{pq} \qquad [\because z = \log_e (p/q)]$$

$$= \frac{1}{pq} \cdot \kappa_{r+1} \qquad \text{[From (***)]}$$

(d) $\dfrac{d\kappa_r}{dz} = \dfrac{d\kappa_r}{dp}\cdot\dfrac{dp}{dz} = \dfrac{d\kappa_r}{dp}\Big/\dfrac{dz}{dp} = \dfrac{d\kappa_r}{dp}\Big/\dfrac{1}{pq} = pq\cdot\dfrac{d\kappa_r}{dp}$

$\therefore \qquad \dfrac{d\kappa_r}{dz} = \kappa_{r+1}$ [c.f. part (c)]

Example 7·23. *If* $b(r;n,p) = \binom{n}{r} p^r q^{n-r}$ *is the binomial probability in the usual notation and if*

$$B(k;n,p) = P(X \le k) = \sum_{r=0}^{k} b(r;n,p),$$

then prove that

$$B(k;n,p) = (n-k)\binom{n}{k}\int_0^q t^{n-k-1}(1-t)^k\,dt;\ q = 1-p$$

Solution. $B(k;n,p) = \sum_{r=0}^{k} b(r;n,p) = \sum_{r=0}^{k}\binom{n}{r}p^r q^{n-r}$

Differentiating w.r. to q and noting that $q = 1-p \Rightarrow \dfrac{dq}{dp} = -1$, we get:

$$\frac{d}{dq}\cdot B(k;n,p) = \sum_{r=0}^{k}\left[\binom{n}{r}\left\{rp^{r-1}(-1)\cdot q^{n-r} + p^r\cdot(n-r)q^{n-r-1}\right\}\right]$$

$$= \sum_{r=0}^{k}\left[\frac{n!\,(-r)}{r!\,(n-r)!}p^{r-1}q^{n-r} + \frac{n!\,(n-r)}{r!\,(n-r)!}p^r q^{n-r-1}\right]$$

$$= \sum_{r=0}^{k}\left[-\frac{n\,(n-1)!}{(r-1)!\,(n-r)!}p^{r-1}q^{n-r} + \frac{n\,(n-1)!}{r!\,(n-r-1)!}p^r q^{n-r-1}\right]$$

$$= \sum_{r=0}^{k}\left[n\cdot\binom{n-1}{r}p^r q^{n-r-1} - n\binom{n-1}{r-1}p^{r-1}q^{n-r}\right]$$

$$= \sum_{r=0}^{k}\left[n\left\{t_r - t_{r-1}\right\}\right] \qquad \ldots(**)$$

$$\text{where } t_r = \binom{n-1}{r}p^r q^{n-r-1} \qquad \ldots(***)$$

$$= n\left[\left(t_0 - t_{-1}\right) + \left(t_1 - t_0\right) + \left(t_2 - t_1\right) + \ldots + \left(t_k - t_{k-1}\right)\right]$$

$$= n\,t_k \qquad [\because t_{-1} = 0, \text{ From } (***)]$$

$$\therefore \frac{d}{dq}\cdot B(k,n,p) = n\binom{n-1}{k}p^k\cdot q^{n-k-1}, \quad p = 1-q$$

On integration, we get

$$B(k;n,p) = n \cdot \binom{n-1}{k} \int_0^q (1-u)^k \cdot u^{n-k-1}\, du.$$

But $n \cdot \binom{n-1}{k} = \dfrac{n \cdot (n-1)!}{k!\,(n-1-k)!} = \dfrac{n!\,(n-k)}{k!\,(n-k)!} = (n-k)\binom{n}{k}$

$$\therefore \qquad B(k;n,p) = (n-k)\binom{n}{k}\int_0^q (1-u)^k \cdot u^{n-k-1}\, du$$

as desired.

Remarks. 1. We further get :

$$\beta(k+1,\, n-k) = \frac{\Gamma(k+1)\,\Gamma(n-k)}{\Gamma(n+1)} = \frac{k!\,(n-k-1)!}{n!}$$

$$\Rightarrow \frac{1}{\beta(k+1,\, n-k)} = \frac{n!}{k!\,(n-k-1)!} = (n-k)\binom{n}{k}$$

Hence the result may be written as :

$$B(k;n,p) = P(X \le k) = \frac{1}{\beta(k+1,\, n-k)}\int_0^q (1-u)^k u^{n-k-1}\, du$$

This result is of great practical utility. It enables us to represent the cumulative Binomial Probabilities (which are generally quite tedious and time consuming to compute) in terms of Incomplete Beta Functions which are tabulated in Karl Pearson's Tables of the Incomplete Beta Functions.

2 Let us now work out the probability :

$$P(X \ge k) = \sum_{r=k}^{n} \binom{n}{r} p^r q^{n-r}$$

Differentiating w.r.t. to p, and proceeding similarly, we shall get :

$$\frac{d}{dp} P(X \ge k) = -n \sum_{r=k}^{n} \left(T_r - T_{r-1}\right) \qquad \text{(Try it)}$$

where $\qquad T_r = \binom{n-1}{r} p^r q^{n-r-1}, \qquad \left(T_n = 0\right)$

$$\therefore \frac{d}{dp} P(X \ge k) = n\, T_{k-1} = n\binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} \quad (\because q = 1-p)$$

On integration, we shall get :

$$P(X \ge k) = n\binom{n-1}{k-1}\int_0^p u^{k-1}(1-u)^{n-k}\, du$$

$$P(X \ge k) = \frac{1}{\beta(k,\, n-k+1)}\int_0^p u^{k-1}(1-u)^{n-k}\, du$$

This is quite an important result and should be committed to memory. We shall use it in *'Order Statistics'*.

This result can be stated as follows :

If $X \sim B(n, p)$ and Y has Beta distribution with parameters k and $n - k + 1$ (c.f. Chapter 8), then

$$P(Y \le p) = P(X \ge k) = 1 - P(X \le k - 1)$$

$$\Rightarrow \qquad F_Y(p) = 1 - F_X(k - 1)$$

EXERCISE 7 (a)

1. (*a*) Describe the probability model from which the binomial distribution can be generated. Hence find the first four central moments.

(*b*) If p is the probability of 'success' at a single trial, obtain the probability of r 'successes' out of n independent trails. Determine the mode of the resulting distribution.

2. (*a*) Define the binomial distribution with parameters p and n, and give a situation in real life where the distribution is likely to be realized. Obtain the moment generating function of the binomial distribution and hence or otherwise obtain the mean, variance, skewness and kurtosis of the distribution.

(*b*) Obtain the Moment Generating Function of the Binomial Distribution. Derive from it the result that the sum of two binomial variates is a binomial variate if the variates are independent and have the same probability of success.

3. The mean and variance of a binomial variate X with parameters n and p are 16 and 8. Find

(*i*) $P(X = 0)$, (*ii*) $P(X = 1)$, (*iii*) $P(X \ge 2)$.

4. For a Binomial distribution the mean is 6 and the standard deviation is $\sqrt{2}$. Write out all the terms of the distribution.

Ans. $n = 9,\ p = 2/3,\ q = 1/3;\ p(r) = (1/3)^9 \cdot \binom{9}{r} 2^r;\ r = 0, 1, 2, \ldots, 9$

5. (*a*) A perfect cube is thrown a large number of times in sets of 8. The occurrence of a 2 or 4 is called a success. In what proportion of the sets would you expect 3 successes.

Ans 27·31%

(*b*) In eight throws of a die, 5 or 6 is considered a success. Find the mean number of successes and the standard deviation. (Ans. 2·66, 1·33)

(*c*) A man tosses a fair coin 10 times. Find the probability that he will have

(*i*) heads on the first five tosses and tails on the next five tosses

(*ii*) heads on tosses 1, 3, 5, 7, 9 and tails on tosses 2, 4, 6, 8, 10.

(*iii*) 5 heads and 5 tails

(*iv*) at least 5 heads

(*v*) not more than 5 heads. **[Madras Univ. B.Sc. (Main Stat) Nov. 1991]**

Ans. (*i*) $(1/2)^{10}$, (*ii*) $(1/2)^{10}$, (*iii*) ${}^{10}C_5\,(1/2)^{10}$

(*iv*) $\sum_{x=5}^{10} {}^{10}C_x \left(\frac{1}{2}\right)^{10}$ (*v*) $\sum_{x=0}^{5} {}^{10}C_x \left(\frac{1}{2}\right)^{10}$

6. (*a*) In 256 sets of twelve tosses of a fair coin, in how many cases may one expect eight heads and four tails ?

(Ans. 31) **(Delhi Univ. B.Sc. Oct. 1992)**

(*b*) In 100 sets of ten tosses of an unbaised coin, in how many cases should we expect

(*i*) Seven heads and three tails, (*ii*) at least seven heads ?

Ans. (*i*) 12, (*ii*) 17

7. (*a*) During war 1 ship out of 9 was sunk on an average in making a certain voyage. What was the probability that exactly 3 out of a convoy of 6 ships would arrive safely ? **(Madras Univ. B.Sc., 1992)**

Ans. $^6C_3\ (8/9)^3\ (1/9)^3$

(*b*) In the long run 3 vessels out of every 100 are sunk. If 10 vessels are out, what is the probability that

(*i*) exactly 6 will arrive safely, and

(*ii*) at least 6 will arrive safely ?

Hint. The probability 'p' that a vessel will arrive safely is

$$P = 97/100 = 0\cdot 97 \text{ and } q = 0\cdot 03$$

The probability that out of 10 vessels, x vessels will arrive safely is

$$p(x) = {}^{10}C_x p^x q^{10-x} = {}^{10}C_x (0\cdot 97)^x (0\cdot 03)^{10-x}$$

(*i*) Required probability $= p(6) = {}^{10}C_6 (0\cdot 97)^6 (0\cdot 03)^x$.

(*ii*) Required probability $= P(X \geq 6)$

8. (*a*) A student takes a true-false examination consisting of 10 questions. He is completely unprepared so he plans to guess each answer. The guesses are to be made at random. For example, he may toss a fair coin and use the outcome to determine his guess.

(*i*) Compute the probability that he guesses correctly at least five times.

(*ii*) Compute the probability that he guesses correctly at least 9 times.

(*iii*) What is the smallest n that the probability of guessing at least n correct answers is less than 1/2. **(Dibrugarh Univ. M.A., 1993)**

Ans. (*i*) 319/512 ; (*ii*) 11/1024 ; (*iii*) 6.

(*b*) A multiple choice test consists of 8 questions and 3 answers to each question, of which only one is correct. If a student answers each question by rolling a balanced die and checking the first answer if he gets 1 or 2, the second answer if he gets 3 or 4, and the third answer if he gets 5 or 6, find the probability of getting:

(*i*) exactly 3 correct answers,

(*ii*) no correct answer,

(*iii*) at least 6 correct answers. **[Gauhati Univ. M.A. (Econ.), 1993]**

9. (*a*) The incidence of occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the probability that out of six workers chosen at random, four or more will suffer from the disease.

Ans. 52/3125

(*b*). (*a*) In a binomial distribution consisting of 5 independent trails, probabilities of 1 and 2 successes are $0\cdot 4096$ and $0\cdot 2048$ respectively. Find the parameter p of the distribution. **(Ans.** $0\cdot 2$**)**

10. (a) With the usual notations, find p for a binomial random variable X if $n = 6$ and if $9P(X = 4) = P(X = 2)$. (Ans. 0·25)

(Mysore Univ. B.Sc. April 1992)

(b) X is a random variable following binomial distribution with mean 2·4 and variance 1·44. Find $P(X \geq 5)$, $P(1 < X \leq 4)$.

11. (a) In a certain town 20% of the population is literate, and assume that 200 investigators take a sample of ten individuals each to see whether they are literate. How many investigators would you expect to report that three people or less are literates in the sample ? **(Shivaji Univ. B.Sc., Oct. 1992)**

(b) A lot contains 1 per cent of defective items. What should be the number (n) of items in a random sample so that the probability of finding at least one defective in it, is at least 0·95 ? (Ans. 68)

12. (a) If on the average rain falls on ten days in every thirty days, find the probability

(i) that rain falls on at least three days of a given week,

(ii) that first three days of a given week will be dry and the remaining wet.

Ans. (i) $\sum\limits_{x=3}^{7} {}^7C_x (1/3)^x (2/3)^{7-x}$, (ii) $(2/3)^3 \cdot (1/3)^4$.

(b) Suppose that weather records show that on the average 5 out of 31 days in October are rainy days. Assuming a binomial distribution with each day of October as an independent trial, find the probability that the next October will have at most three rainy days.

Ans. 0·2403

13. The probability of a man hitting a target is 1/4. (i) If he fires 7 times, what is the probability p of his hitting the target at least twice ? (ii) How many times must he fire so that the probability of his hitting the target at least once is greater than 2/3 ? [Ans. (i) 4547/8192, (ii) 4]

Hint. (ii) $p = \frac{1}{4}$, $q = \frac{3}{4}$ · We want n such that

$$1 - q^n > \frac{2}{3} \Rightarrow q^n < \frac{1}{3} \Rightarrow \left(\frac{3}{4}\right)^n < \frac{1}{3} \Rightarrow n = 4$$

14. (a) The probability of a man hitting a target is 1/3. How many times must be fire so that the probability of hitting the target at least once is more than 90%. Ans. 6. **(Shivaji Univ. B.Sc., 1991)**

(b) Eight mice are selected at random and they are divided into two groups of 4 each. Each mouse in group A is given a dose of certain poison 'a' which is expected to kill one in four; each mouse in group B is given a dose of certain poison 'b' which is expected to kill one or two. Show that nevertheless, there may be fewer deaths in group A and find the probability of this happening.

Ans. 525/4096

15 (a) A card is drawn and replaced in an ordinary deck of 52 cards. How many times must a card be drawn so that (i) there is at least an even chance of drawing a heart, (ii) the probability of drawing a heart is greater than 3/4 ?

Ans. (i) 3, (ii) 5

(*b*) Five coins are tossed. What is the variance of the number of heads per toss of the five coins:

(*i*) if each coins is unbiased,

(*ii*) if the probability of a head appearing is 0·75 for each coin, and

(*iii*) if four coins are unbaised and for the fifth the probability of a head appearing is 0·75 ?

Hint (*iii*) Use generating function. [See Ex. 7·17 (*iii*)]

16. An owner of a small hotel with five rooms is considering buying television sets to rent to room occupants. He expects that about half of his customers would be willing to rent sets, and finally he buys three sets. Assuming 100% occupancy at all times :

(*i*) . What fraction of the evenings will there be more request than T.V. sets?

(*ii*) What is the probability that a customer who requests a television set will receive one?

(*iii*) If the owner's cost per set per day is C, what rent R must he charge in order to break even (neither gain nor lose) in the long run ?

Hint. (*i*) Let the random variable X denote the daily number of requests. Then required probability is

$$P(X \geq 4) = P(X = 4) + P(X = 5) = \binom{5}{4}\left(\frac{1}{2}\right)^5 + \binom{5}{5}\left(\frac{1}{2}\right)^5$$

(*ii*) The customer can get a T.V. in the following mutually exclusive ways,

(*a*) There are no other requests that night.

(*b*) There is one other request.

(*c*) There are two other requests.

(*d*) There are three other requests and his request precedes at least one of them.

(*e*) There are four other requests, and his request preceedes at least two of them.

The probability of the desired event

$$= (0\cdot 5)^4 \left\{1 + {}^4C_1 + {}^4C_2 + \frac{3}{4}\cdot {}^4C_3 + \frac{3}{5}\,{}^4C_4\right\}$$

(*iii*) Mean revenue

$$= (0\cdot5)^5\cdot 0 + {}^5C_1(0\cdot5)^5 R + {}^5C_2(0\cdot5)^5\, 2R + \left\{{}^5C_3(0\cdot5)^5 + {}^5C_4(0\cdot5)^5 + {}^5C_5(0\cdot5)^5\right\} 3R$$

$$= \frac{73}{32} R$$

The break-even rental is the value of R for which

$$\frac{73}{32} R = 3C \quad \Rightarrow \quad R = 1\cdot 315\, C$$

17. A manufacturer claims that at most 10 per cent of his product is defective. To test this claim, 18 units are inspected and his claim is accepted if among these 18 units, at most 2 are defective. Find the probability that the manufacturer's claim will be accepted if the actual probability that a unit is defective is

(*a*) 0·05 (*b*) 0·10 (*c*) 0·15 and (*d*) 0·20.

Ans. (*a*) 0·9410 (*b*) 0·9326 (*c*) 0·4445 (*d*) 0·2715

18. (*a*) A set of 8 symmetrical coins was tossed 256 times and the frequencies of throws observed were as follows :

Number of heads :	0	1	2	3	4	5	6	7	8
Frequency of throws:	2	6	24	63	64	50	36	10	1

Fit a binomial distribution and find mean and standard deviation of fitted distribution.

(*b*) A set of 6 similar coins is tossed 640 times with the following results:

Number of heads :	0	1	2	3	4	5	6
Frequency :	7	64	140	210	132	75	12

Calculate the binomial frequencies on the assumption that the coins are symmetrical.

19. (*a*) The following data due to Weldon shows the results of throwing 12 dice 4096 times, a throw of 4, 5 or 6 being called a success (x).

x :	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
f :	—	7	60	198	430	731	948	847	536	257	71	11	—	4096

Fit the binomial distribution and calculate the expected frequencies. Compare the actual mean and S.D. with those of the expected ones for the distribution.

Ans. Expected freq. : 1, 12, 66, 220 495 792, 924, 792, 495, 220, 66, 12, 0; mean = 6, variance = 1·71.

(*b*) In 103 litters of 4 mice, the number of litters which contained 0, 1, 2, 3, 4 females are recorded below :

Number of female mice	0	1	2	3	4	Total
Number of litters	8	32	34	24	5	103

(*i*) If the chance of obtaining a female in a single trial is assumed constant, estimate the constant but unknown probability.

(*ii*) If the size of the litter 4 had not been given, how could it be estimated from the data ?

20. X is random variable distributed according to the Binomial law :

$$b(x\,;n,p) = \binom{n}{x} p^x q^{n-x}\,;\; x = 0, 1, 2, \ldots, n$$

Obtain the recurrence formula :

$$b(x+1\,;n,p) = \frac{n-x}{x+1}\cdot\frac{p}{q}\cdot b(x\,;n,p)$$

Use this as a reduction formula and get the theoretical frequencies when an unbaised coin is tossed 8 times and the experiment is repeated 256 times.

(Madras Univ. B. Sc. April 1992)

21. (*a*) By differentiating the following identity with respect to p and then multiplying by p,

$$\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (q+p)^n, q = 1-p$$

prove that $\mu_1' = np$ and $\mu_2 = npq$.

22. (*a*) Let $X \sim b(x; n, p)$ and r be a non-negative integer. If the rth moment about the origin is denoted by $\mu_r' = E(X^r)$, prove that

$$\mu_{r+1}' = np\,\mu_r' + p(1-p)\frac{d\mu_r'}{dp}$$

[Delhi Univ. B.Sc. (Hons. Subs.), 1993, '88]

(*b*) Show that for the binomial distribution $B(n, p)$,

$$\mu_{r+1} = pq\left(nr\,\mu_{r-1} + \frac{d}{dp}\mu_r\right), \quad p + q = 1,$$

where symbols have their usual meanings.

[Delhi Univ. B.Sc. (Stat. Hons), 1989]

(c) If $X \sim B(n, p)$, obtain the recurrence relation for its central moments and hence find values of β_1 and β_2.

[Calcutta Univ. B.Sc. (Hons.), 1992]

23. (*a*) The following results were obtained when 100 batches of seeds were allowed to germinate on damp filter paper in a laboratory :

$$\beta_1 = \frac{1}{15} \text{ and } \beta_2 = \frac{89}{30}$$

Determine the binomial distribution and calculate the frequency for $X = 8$, considering $p > q$.

Hint. We have $\beta_1 = \dfrac{(q-p)^2}{npq} = \dfrac{1}{15}$ ...(*i*)

and $\beta_2 = 3 + \dfrac{1 - 6pq}{npq} = \dfrac{89}{30}$ ...(ii)

From (*i*) and (*ii*), we can find the value of n, p and q

(*b*) Between a Binomial distribution with $n = 5$ and $p = \frac{1}{2}$ and a distribution with frequency function

$$f(x) = 6x(1-x), \; 0 \le x \le 1 \;;$$

determine which is more skewed.

24. (*a*) $x = r$ is the unique mode of Binomial Distribution having mean np and variance $np(1-p)$. Show that

$$(n+1)p - 1 < r < (n+1)p$$

Find the mode of the binomial distribution with $p = \frac{1}{2}$ and $n = 7$.

[Delhi Univ. B.Sc. (Stat. Hons.) 1991, '84]

Ans. 4, 3 (Bimodal).

(*b*) Show that if np be a whole number, the mean of the binomial distribution coincides with the greatest term.

(*c*) Compute the mode of a binomial distribution $b\left(7, \frac{1}{2}\right)$.

[Delhi Univ. B.Sc. (Maths. Hons.), 1989]

Ans. 1, 2 (Bimodal).

(*d*) Define Bernoulli trials and state the binomial law of probability. Find the bounds for the most probable number of successes in a sequence of n Bernoulli trials.

One workers can manufacture 120 articles during a shift, another worker 140 articles, the probabilities of the articles being of a high quality are 0·94 and 0·80 respectively. Determine the most probable number of high quality articles manufactured by each worker. **[Calcutta Univ. B.Sc. (Maths. Hons.), 1988]**

25. Show that if two symmetrical binomial distributions ($p = q = \frac{1}{2}$) of degree n (and of the same number of observations) are so superimposed that the rth term of one coincides with the $(r + 1)$th term of the other, the distribution formed by adding superimposed terms is a symmetrical binomial of degree $(n + 1)$. **[Bhagalpur Univ. B.Sc., 1993]**

26. (*a*) Let X denote a binomially distributed random variable. Show that

$$E\left(\frac{X-np}{\sqrt{npq}}\right) = 0,\ E\left(\frac{X-np}{\sqrt{npq}}\right)^2 = 1, \text{ and}$$

$$E\left[\exp.\left\{t\left(\frac{X-np}{\sqrt{npq}}\right)\right\}\right] = \left[(1-p)\exp\left\{-t\sqrt{\left(\frac{p}{nq}\right)}\right\} + p\exp\left\{t\sqrt{\left(\frac{q}{np}\right)}\right\}\right]$$

(*b*) Obtain the characteristic function of the standard binomial variate $(X - np)/\sqrt{npq}$, where X is the number of successes obtained in n independent trials, each with constant probability p of success, $q = 1 - p$. Obtain the limit of this function as $n \to \infty$. **[Delhi Univ. B.Sc.(Maths. Hons.), 1991]**

(*c*) If $X \sim B(n, p)$, prove that

$$\kappa_{r+1} = pq \cdot \frac{d}{dp}(\kappa_r),$$

where κ_r is the rth cumulant.

Hence deduce the values of κ_2 and κ_3.

[Delhi Univ. B.Sc. (Stat. Hons.), 1991, '87]

27. (*a*) If X and Y are two independent identically distributed binomial variates, obtain the probability that the absolute difference $| X - Y |$ equals a given value, say r.

(*b*) (*i*) If X and Y are independent binomial variates, with parameters p_1 and p_2 and indices n_1 and n_2 respectively, obtain the probability that $X + Y$ equals 'r'.

(*ii*) In the above if $p_1 = p_2$, what is the distribution of $X + Y$?

[Poona Univ. B.Sc., 1988]

(*c*) If X and Y are two independent binomial variates with parameters $n_1 = 6,\ p = 1/2$ and $n_2 = 4,\ p = 1/2$ respectively, evaluate,

(*i*) $P(X + Y = r)$, (*ii*) $P(X + Y \geq 3)$

(Gujarat Univ. B. Sc. Oct. 1992)

Hint $X + Y \sim B(6 + 4,\ 1/2) = B(10, 1/2)$

Ans. (*i*) $P(X + Y = r) = p(r) = {}^{10}C_r\ (1/2)^r;\ r = 0, 1, ..., 10$

(*ii*) $P(X + Y \geq 3) = 1 - [p(0) + p(1) + p(2)] = 0{\cdot}945$

(*d*) If X and Y are two independent binomial variates with parameters $(n_1 = 3, p = 0{\cdot}4)$ and $(n_2 = 4,\ p = 0{\cdot}4)$ respectively, find :

(*i*) $P(X = Y)$, (*ii*) $P(X + Y \leq 2)$, (*iii*) $P(X = 3 \mid X + Y = 4)$

Hint. $X + Y \sim B(3 + 4, 0{\cdot}4) = B(7, 0{\cdot}4)$

$$(i)\ P(X = Y) = \sum_{r=0}^{3} P(X = r \cap Y = r) = \sum_{r=0}^{3} P(X = r)\ P(Y = r) = 0{\cdot}2871$$

$$(ii)\ P(X + Y \leq 2) = \sum_{r=0}^{2} \binom{7}{r} (0{\cdot}4)^r (0{\cdot}6)^{7-r} = 0{\cdot}420$$

$$(iii)\ P(X = 3 \mid X + Y = 4) = \frac{P(X = 3 \cap X + Y = 4)}{P(X + Y = 4)} = \frac{P(X = 3 \cap Y = 1)}{P(X + Y = 4)} = 0{\cdot}1141$$

28. (*a*) Obtain the moment generating function of Binomial distribution with $n = 7$ and $p = 0{\cdot}6$. Find the first three moments of the distribution.

[Poona Univ. B. Sc. 1992)

Ans. $(0{\cdot}4 + 0{\cdot}6\,e^t)^7$: mean $= 4{\cdot}2,\ \mu_2 = 1{\cdot}68,\ \mu_3 = -0{\cdot}336$.

(*b*) Suppose that the m.g.f. of a random variable X is of the form

$$M_X(t) = (0{\cdot}4\,e^t + 0{\cdot}6)^8$$

What is the m.g.f. of the random variable $Y = 3X + 2$? Evaluate $E(X)$.

Ans. $E(X) = 3{\cdot}2,\quad M_Y(t) = e^{2t}\,(0{\cdot}6 + 0{\cdot}4\,e^{3t})^8$

(*c*) Obtain the moment generating function of the binomial distribution. Hence or otherwise obtain the mean, variance and skewness of the distribution.

29. Show that the factorial moment generating function $w(t)$ of the binomial distribution $b(x; n, p)$ is $(1 + pt)^n$. Hence or otherwise show that

$$\mu_{(r)}' = n^{(r)}\,p^r$$

Hint. Factorial moment generating function $\omega(t)$ is defined as

$$\omega(t) = E(1 + t)^X = \sum_x (1 + t)^x p(x) = \sum_x {}^nC_x \{p(1 + t)\}^x q^{n-x}$$

$$\mu_{(r)}' = \text{co efficient of } \frac{t^r}{r!} \text{ in } \omega(t) = {}^nC_r\, r!\, p^r = n^{(r)} p^r$$

30. Show that

(*i*) $b(n, p; k) = b(n,\ 1 - p;\ n - k)$

(*ii*) $\sum_{k=r}^{n} b(n, p; k) = 1 - \sum_{k=n-r+1}^{n} b(n,\ 1 - p; k)$

(*iii*) $b(n + 1, p; k) = p\,.\,b(n, p; k - 1) + q\,.\,b(n, p; k)$

Hint. (*i*) $b(n, 1-p; n-k) = \binom{n}{n-k}(1-p)^{n-k} p^{n-(n-k)}$

(*ii*) $\sum_{k=r}^{n} b(n, p; k) = \sum_{k=r}^{n} b(n, 1-p; n-k) = \sum_{k=0}^{n-r} b(n, 1-p; k)$

31. For a binomial distribution, let

$$F_n(y) = \sum_{x=0}^{y} \binom{n}{x} p^x q^{n-x},$$

where $q = 1-p$,
prove that

(*i*) $F_{n+1}(y) = p\, F_n(y-1) + q\, F_n(y)$

(*ii*) $\text{Cov}(X, n-X) = -npq$ **(Bombay Univ. B.Sc., April 1990)**

32. (*a*) Random variable X follows binomial distribution with parameters $n = 40$ and $p = \frac{1}{4}$ · Use Chebychev's inequality to find bounds for

(*i*) $P[\,|X-10| < 8]$; (*ii*) $P[\,|X-10| > 10]$

Compare these values with the actual values (**Hint** : Use Normal approximation for the Binomial). **(Madras Univ. B.Sc. (Main Stat.), 1988)**

Ans. (*i*) 113/128 (lower bound), (*ii*) 0.075 (upper bound).

(*b*) X follows binomial distribution with $n = 40$, $p = \frac{1}{2}$ · Use Chebychev's lemma to

(*i*) find k such that
$P\{\,|X-20| > 10k\} \le 0{\cdot}25$, and

(*ii*) obtain a lower limit for $P\{|X-20| \le 5\}$.

[Delhi Univ. B.Sc. (Maths. Hons.), 1984]

Ans. (*i*) $2\sqrt{10}$, (*ii*) 3/5

(*c*) How many trials must be made of an event with binomial probability of success $\frac{1}{2}$ in each trial, in order to be assured with probability of at least 0·9 that the relative frequency of success will be between 0·48 and 0·52 ? (**Ans.** 6250)

Hint. Use Chebychev's Inequality.

33. (*a*) Show that if a coin is tossed n times, the probability of not more than k heads is :

$$\left[\binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{k}\right]\left(\frac{1}{2}\right)^n$$

[South Gujarat Univ. B.Sc., 1988]

(*b*) If X has binomial distribution with paramctes n and p, then prove that
$P[X \text{ is even}] = \frac{1}{2}[1 + (q-p)^n]$. **[Delhi Univ. B.Sc. (Stat. Hons.), 1988]**

34. If the probability of hitting a target is 1/5 and if 10 shots are fired, what is the conditional probability of the target being hit at least twice assuming that at least one hit is already scored ?

[Nagpur Univ. B.Sc., 1988, '93]

Hint. Let X donote the number of times a target is hit when 10 shots are fired. Then $X \sim B(10, 0\cdot 2)$. The required probability is :

$$P(X \geq 2 \mid X \geq 1) = \frac{P[(X \geq 2) \cap (X \geq 1)]}{P(X \geq 1)} = \frac{P(X \geq 2)}{P(X \geq 1)}$$

$$= \frac{1 - [P(X = 0) + P(X = 1)]}{1 - [P(X = 0)]} = \frac{0\cdot 625}{0\cdot 893} = 0\cdot 6999$$

35. (a) Let X be a $B(2, p)$ and Y be a $B(4, p)$. If $P(X \geq 1) = 5/9$, find $P(Y \geq 1)$ **[Kerala Univ. B. Sc., 1989]**

Hint. $P(X \geq 1) = 1 - P(X = 0) = 1 - q^2 = 5/9 \Rightarrow q = 2/3, p = 1/3.$

$P(Y \geq 1) = 1 - P(Y = 0) = 1 - q^4 = 65/81.$

36. Let B denote the number of boys in a family with five children. If p denotes the probability that a boy is there in a family, find the least value of p such that

$P(B = 0) > P(B = 1)$ **(Shivaji Univ. B. Sc., 1990)**

Ans. $q^5 > 5pq^4 \Rightarrow q > 5p \Rightarrow p < \frac{1}{6}.$

37. (a) Suppose $X \sim B(n, p)$. with $E(X) = 5$, Var $(X) = 4$. Find n and p. (Ans. $n = 25$, $p = 1/5$)

(b) Let $X \sim B(n, p)$. For what p is variance (X) maximised if we assume n is fixed.

Ans. Var $X = npq = n(p - p^2) = f(p)$, (say) ; $f'(p) = 0$, $f''(p) < 0$; $p = 1/2 = q$

38. (a) $X \sim B(n = 100, p = 0\cdot 1)$. Find $P(X \leq \mu_x - 3\sigma_x)$

Ans. $\mu = 10$, $\sigma = 3$, $P(X \leq \mu_x - 3\sigma_x) = P(X \leq 1) = 10{\cdot}9 \times (0{\cdot}9)^{99}$

(b) If $X \sim B(25, 0\cdot 2)$, find $P(X < \mu_x - 2\sigma_x)$

[Delhi Univ. B.A. (Stat. Hons.) Spl. Course 1989]

39. For one half of n events, the chance of success is p, and the chance of failure is q, whilst for the other half the chance of success is q, and the chance of failure is p. Show that the S.D. of the number of successes is the same as if the chance of success were p in all the cases i.e. $\sqrt{npq}$, but that the mean of the number of successes is $n/2$ and not np. **(Delhi Univ. B.A. 1992)**

Hint. $X \sim B(n/2, p)$ and $Y \sim B(n/2, q)$ are independent. Let $Z = X + Y$. Now prove that Var $(Z) = npq$ and $E(Z) = n/2$.

40. The discrete density of X is given by $f_X(x) = x/3$, for $x = 1, 2$ and $f_{Y|X}(y \mid x)$ is binomial with parameters x and $\frac{1}{2}$ i.e.,

$$F_{Y|X}(y \mid x) = P(Y = y \mid X = x) = \binom{x}{y} \cdot \left(\tfrac{1}{2}\right)^x ;$$

for $y = 0, 1, \ldots, x$ and $x = 1, 2$.

(a) Find $E(X)$ and Var (X) ; (b) Find $E(Y)$

(c) Find the joint distribution of X and Y.

Hint. Proceed as in Example 7·21.

Ans. (a) $E(X) = 5/3$, $\text{Var}(X) = 2/9$, (b) $E(Y) = 5/6$.

(c) $f(x,y) = \binom{x}{y} \cdot \left(\frac{x}{3}\right) \cdot \left(\frac{1}{2}\right)^x$; $n = 1, 2, ;$ $y = 0, 1, \ldots, x.$

41. Two dice are thrown n times. Let X denote the number of throws in which the number on the first dice exceeds the number on the second dice. What is the distribution of X?

Ans. $X \sim B(n, p = 15/36)$

Hint. p is the probability that the number on the first dice exceeds the number on the second dice in a throw of two dice.

42. Let $X_1 \sim B(n, p_1)$ and $X_2 \sim B(n, p_2)$.

If $p_1 < p_2$, prove that:

$$P(X_1 \le k) \ge P(X_2 \le k) \text{ for } k = 0, 1, \ldots, n.$$

Hint. Use Example 7·23.

43. If $X \sim B(n, p)$, show that

$$P(X \le k) = \lambda \int_{p/q}^{\infty} \frac{y^k}{(1+y)^{n+1}}\, dy$$

where $$\lambda^{-1} = \int_0^{\infty} \frac{y^k}{(1+y)^{n+1}}\, dy = \beta(k+1, n-k)$$

Hint. $\dfrac{d}{dq} P(X \le k) = n\binom{n-1}{k} \cdot p^k \cdot q^{n-k-1} = A_k$, (say)

[See Example 7·23]

Find $\dfrac{d}{dq}(\text{RHS}) = \lambda \cdot \dfrac{d}{dq}\left(\displaystyle\int_{p/q}^{\infty} \frac{y^k}{(1+y)^{n+1}}\, dy\right) = \lambda \dfrac{(p/q)^k}{[1+(p/q)]^{n+1}}\left(\dfrac{1}{q^2}\right)$

$$= \frac{1}{\beta(k+1, n-k)} \cdot p^k \cdot q^{n-k-1} = A_k$$

(On simplification)

44. If X $B(n, p)$ and Y has beta distribution with parameters k and $n-k+1$, (See Chapter 8), then prove that

$P(Y \le p) = P(X \ge k)$ i.e., $F_Y(p) = 1 - F_X(k-1)$

45. If a fair coin is tossed an even number $2n$ times, show that the probability of obtaining more heads than tails is

$$\frac{1}{2}\left\{1 - {}^{2n}C_n \left(\frac{1}{2}\right)^{2n}\right\}$$

Hint. X : No. of heads; Y = No. of tails; No. of trials = $2n$

$P(X > Y) + P(X < Y) + P(X = Y) = 1$

$\Rightarrow$ $$2P(X > Y) = 1 - P(X = Y)$$

[$\because$ By symmetry, $p = q = \frac{1}{2} \Rightarrow P(X > Y) = P(X < Y)$]

$$= 1 - {}^{2n}C_n p^n \cdot q^n = 1 - {}^{2n}C_n \left(\tfrac{1}{2}\right)^{2n}$$

$$\Rightarrow \qquad P(X > Y) = \frac{1}{2}\left[1 - {}^{2n}C_n \left(\tfrac{1}{2}\right)^{2n}\right]$$

7·3·0. Poisson Distribution (as a limiting case of Binomial Distribution). Poisson distribution was discovered by the French mathematician and physicist Simeon Denis Poisson (1781—1840) who published it in 1837. Poisson distribution is a limiting case of the binomial distribution under the following conditions:

(*i*) n, the number of trials is indefinitely large, *i.e.*, $n \to \infty$.

(*ii*) p, the constant probability of success for each trial is indefinitely small, *i.e.*, $p \to 0$.

(*iii*) $np = \lambda$, (say), is finite. Thus $p = \lambda/n$, $q = 1 - \lambda/n$, where λ is a positive real number.

The probability of x successes in a series of n independent trials is

$$b(x\,; n, p) = \binom{n}{x} p^x q^{n-x}\,;\ x = 0, 1, 2, \ldots, n \qquad \ldots(*)$$

We want the limiting form of (*) under the above conditions. Hence

$$\lim_{n\to\infty} b(x\,; n, p) = \lim_{n\to\infty} \frac{n\,!}{x\,!\,(n-x)\,!}\left(\frac{\lambda}{n}\right)^x \cdot \left[1 - \frac{\lambda}{n}\right]^{n-x}$$

Using Stirling's approximation for $n\,!$ as $n \to \infty$ *viz.*,

$$\lim_{n\to\infty} n\,! \approx \sqrt{2\pi}\, e^{-n}\, n^{n+(1/2)}, \text{ we get}$$

$$\lim_{n\to\infty} b(x; n, p) = \lim_{n\to\infty}\left[\frac{\sqrt{2\pi}\, e^{-n}\cdot n^{n+(1/2)}}{x\,!\,\sqrt{2\pi}\, e^{-(n-x)}\cdot (n-x)^{n-x+(1/2)}}\right]\left(\frac{\lambda}{n}\right)^x\left[1 - \frac{\lambda}{n}\right]^{n-x}$$

$$= \frac{\lambda^x}{e^x \cdot x\,!}\cdot \lim_{n\to\infty} \frac{n^{n-x+(1/2)}}{(n-x)^{n-x+(1/2)}}\cdot\left[1 - \frac{\lambda}{n}\right]^{n-x}$$

$$= \frac{\lambda^x}{e^x\, x\,!} \lim_{n\to\infty} \frac{\left(1 - \frac{\lambda}{n}\right)^{n-x}}{\left(1 - \frac{x}{n}\right)^{n-x+(1/2)}}$$

$$= \frac{\lambda^x}{e^x\, x\,!}\cdot \frac{\lim\limits_{n\to\infty}\left[1 - \frac{\lambda}{n}\right]^n \cdot \lim\limits_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{-x}}{\lim\limits_{n\to\infty}\left[1 - \frac{x}{n}\right]^n \lim\limits_{n\to\infty}\left[1 - \frac{x}{n}\right]^{-x+(1/2)}}$$

But we know that

and $$\left.\begin{array}{l} \lim\limits_{n \to \infty} \left(1 - \dfrac{\lambda}{n}\right)^{n} = e^{-\lambda}, \\[2ex] \lim\limits_{n \to \infty} \left[1 - \dfrac{\lambda}{n}\right]^{\alpha} = 1, \ \alpha \text{ is not a function of } n \end{array}\right\} \quad ...(**)$$

Therefore

$$\lim_{n \to \infty} b(x\,; n, p) = \frac{\lambda^x}{e^x \cdot x\,!} \cdot \frac{e^{-\lambda} \cdot 1}{e^{-x} \cdot 1} = \frac{e^{-\lambda} \cdot \lambda^x}{x\,!}\,; x = 0, 1, 2, ..., \infty\,;$$

[Using (**)]

which is the required probability function of the Poisson distribution. 'λ' is known as the parameter of Poisson distribution.

Aliter. Poisson distribution can also be derived without using Stirling's approximation as follows :

$$b(x\,; n, p) = \binom{n}{x} p^x (1 - p)^{n-x} = \binom{n}{x} \left[\frac{p}{1-p}\right]^x (1 - p)^n$$

$$= \frac{n(n-1)(n-2)\ldots(n-x+1)}{x\,!} \cdot \frac{\left(\dfrac{\lambda}{n}\right)^x}{\left[1 - \dfrac{\lambda}{n}\right]^x} \left[1 - \frac{\lambda}{n}\right]^n$$

$$= \frac{\left[1 - \dfrac{1}{n}\right]\left[1 - \dfrac{2}{n}\right] \ldots \left[1 - \dfrac{x-1}{n}\right]}{x\,! \left[1 - \dfrac{\lambda}{n}\right]^x} \lambda^x \left[1 - \frac{\lambda}{n}\right]^n$$

$$\therefore \quad \lim_{n \to \infty} b(x\,; n, p) = \frac{e^{-\lambda} \lambda^x}{x\,!}\,; \ x = 0, 1, 2, \ldots$$

[From (**)]

Definition. *A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by*

$$p(x, \lambda) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x\,!}\,; \ x = 0, 1, 2, \ldots\,; \lambda > 0$$
$$= 0, \ \textit{otherwise} \qquad ...(7\cdot 14)$$

Here λ is known as the parameter of the distribution.

We shall use the notation $X \sim P(\lambda)$ to denote that X is a Poisson variate with paramter λ.

Remarks 1. It should be noted that

$$\sum_{x=0}^{\infty} P(X = x) = e^{-\lambda} \sum_{x=0}^{\infty} \lambda^x/x\,! = e^{-\lambda}\, e^{\lambda} = 1$$

2. The corresponding distribution function is:

$$F(x) = P(X \leq x) = \sum_{r=0}^{x} p(r) = e^{-\lambda} \sum_{r=0}^{x} \lambda^r/r\,!\ ; x = 0, 1, 2, \ldots.$$

3. Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials (unlike that in binomial) of an experiment but which occur at random points of time and space wherein our interest lies only in the number of occurrences of the event, not in its non-occurrences.

4. Following are some instances where Poisson distribution may be successfully employed·

(1) Number of deaths from a disease (not in the form of an epidemic) such as heart attack or cancer or due to snake bite.

(2) Number of suicides reported in a particular city.

(3) The number of defective material in a packing manufactured by a good concern.

(4) Number of faulty blades in a packet of 100.

(5) Number of air accidents in some unit of time.

(6) Number of printing mistakes at each page of the book.

(7) Number of telephone calls received at a particular telephone exchange in some unit of time or connections to wrong numbers in a telephone exchange.

(8) Number of cars passing a crossing per minute during the busy hours of a day.

(9) The number of fragments received by a surface area 't' from a fragment atom bomb.

(10) The emission of radioactive (alpha) particles.

7·3·1. The Poisson Process. The Poisson distribution may also be obtained independently (*i.e.*, without considering it as a limiting form of the Binomial distribution) as follows :

Let X_t be the number of telephone calls received in time interval 't' on a telephone switch board. Consider the following experimental conditions :

(1) The probability of getting a call in small time interval $(t, t + dt)$ is $\lambda\, dt$, where λ is a positive constant and dt denotes a small increment in time 't'.

(2) The probability of getting more than one call in this time interval is very small, *i.e.*, is of the order of $(dt)^2$ *i.e.*, $0\,[(dt)^2]$ such that

$$\lim_{dt \to 0} \frac{0\,(dt)^2}{dt} = 0$$

(3) The probability of any particular call in the time interval $(t, t + dt)$ is independent of the actual time t and also of all previous calls.

Under these conditions it can be shown that the probability of getting x calls in time 't', say, $P_x(t)$ is given by

$$P_x(t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}; \; x = 0, 1, 2, ..., \infty$$

which is a Poisson distribution with parameter λt.

Proof. Let $P_x(t) = P$ {of getting x calls in a time interval of length 't'}. Also P {of at least one call during $(t, t + dt)$} $= \lambda\, dt + 0\,[(dt)^2]$

and P {of more than one call during $(t, t + dt)$} $= 0\,[(dt)^2]$.

The event of getting exactly x calls in time $t + dt$ can materialise in the following two mutually exclusive ways :

(*i*) x calls in $(0, t)$ and none during $(t, t + dt)$ and the probability of this event is $P_x(t)\left\{1 - [(\lambda\, dt + 0\,(dt)^2]\right\}$,

(*ii*) exactly $(x - 1)$ calls during $(0, t)$ and one call in $(t, t + dt)$ and the probability of this event is $P_{x-1}(t) \cdot (\lambda\, dt)$.

Hence by the addition theorem of probability, we get

$$P_x(t + dt) = P_x(t)\left\{1 - \lambda\, dt - 0\,(dt)^2\right\} + P_{x-1}(t)\,\lambda\, dt.$$
$$= P_x(t)(1 - \lambda\, dt) + P_{x-1}(t)\,\lambda\, dt + 0\,(dt)^2 P_x(t) \quad ...(1)$$
$$\Rightarrow \frac{P_x(t + dt) - P_x(t)}{dt} = -\lambda P_x(t) + \lambda P_{x-1}(t) + \frac{0\,(dt)^2}{dt} P_x(t)$$

Proceeding to the limit as $dt \to 0$, we get

$$\lim_{dt \to 0} \frac{P_x(t + dt) - P_x(t)}{dt} = -\lambda P_x(t) + \lambda P_{x-1}(t)$$
$$\therefore \qquad P_x'(t) = -\lambda P_x(t) + \lambda P_{x-1}(t), x \geq 1 \quad ...(2)$$

where ($'$) denotes differentiation w.r. to 't'.

For $x = 0, P_{x-1}(t) = P_{-1}(t) = P\left\{(-1) \text{ calls in time 't'}\right\} = 0$

Hence from (1), we get

$$P_0(t + dt) = P_0(t)\left\{1 - \lambda\, dt\right\} + 0\,(dt)^2$$

which on taking the limit $dt \to 0$, gives

$$P_0'(t) = -\lambda P_0(t) \Rightarrow \frac{P_0'(t)}{P_0(t)} = -\lambda$$

Integrating w.r. to. 't', we get

$$\log P_0(t) = -\lambda t + C,$$

where C is an arbitrary constant to be determined from the condition

$$P_0(0) = 1$$

Hence $\qquad \log 1 = C \Rightarrow C = 0$

$\therefore \qquad \log P_0(t) = -\lambda t \Rightarrow P_0(t) = e^{-\lambda t}$

Substituting this value of $P_0(t)$ in (2), we get, with $x = 1$

$$P_1'(t) = -\lambda P_1(t) + \lambda e^{-\lambda t}$$
$$\Rightarrow \qquad P_1'(t) + \lambda P_1(t) = \lambda e^{-\lambda t}$$

This is an ordinary linear differential equation whose integrating factor is $e^{\lambda t}$. Hence its solution is

$$e^{\lambda t} P_1(t) = \lambda \int e^{\lambda t} e^{-\lambda t} \, dt + C_1 = \lambda t + C_1,$$

where C_1 is an arbitrary constant to be determined from $P_1(0) = 0$, which gives $C_1 = 0$.

$$\therefore \qquad P_1(t) = e^{-\lambda t} \lambda t$$

Again substituting this in (2) with $x = 2$, we get

$$P_2'(t) + \lambda P_2(t) = \lambda e^{-\lambda t} \lambda t$$

Integrating factor of this equation is $e^{\lambda t}$ and its solution is

$$P_2(t) e^{\lambda t} = \lambda^2 \int t e^{-\lambda t} e^{\lambda t} \, dt + C_2 = \frac{\lambda^2 t^2}{2} + C_2$$

where C_2 is an arbitrary constant to be determined from $P_2(0) = 0$, which gives $C_2 = 0$. Hence

$$P_2(t) = e^{-\lambda t} \frac{(\lambda t)^2}{2}$$

Proceeding similarly step by step, we shall get

$$P_x(t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!} ; \; x = 0, 1, 2, \ldots, \infty.$$

7·3·2. Moments of the Poisson Distribution

$$\mu_1' = E(X) = \sum_{x=0}^{\infty} x \, p(x, \lambda)$$

$$= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda} \lambda^x}{x!} = \lambda e^{-\lambda} \left[\sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right]$$

$$= \lambda e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots \right) = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

Hence the mean of the Poisson distribution is λ.

$$\mu_2' = E(X^2) = \sum_{x=0}^{\infty} x^2 p(x, \lambda) = \sum_{x=0}^{\infty} \{ x(x-1) + x \} \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} x(x-1) \frac{\lambda^x}{x!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= \lambda^2 e^{-\lambda} \left[\sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} \right] + \lambda = \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda$$

$$\mu_3' = E(X^3) = \sum_{x=0}^{\infty} x^3 p(x, \lambda)$$

$$= \sum_{x=0}^{\infty} \{ x(x-1)(x-2) + 3x(x-1) + x \} \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} x(x-1)(x-2) \frac{e^{-\lambda} \lambda^x}{x!} + 3 \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda} \lambda^x}{x!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= e^{-\lambda}\lambda^3\left[\sum_{x=3}^{\infty}\frac{\lambda^{x-3}}{(x-3)\,!}\right] + 3e^{-\lambda}\lambda^2\left[\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)\,!}\right] + \lambda$$

$$= e^{-\lambda}\lambda^3 e^{\lambda} + 3e^{-\lambda}\lambda^2 e^{\lambda} + \lambda = \lambda^3 + 3\lambda^2 + \lambda$$

$$\mu_4' = E(X^4) = \sum_{x=0}^{\infty} x^4 \cdot p(x; \lambda)$$

$$= \sum_{x=0}^{\infty}\{x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x\}\frac{e^{-\lambda}\lambda^x}{x\,!}$$

$$= e^{-\lambda}\lambda^4\left[\sum_{x=4}^{\infty}\frac{\lambda^{x-4}}{(x-4)\,!}\right] + 6e^{-\lambda}\lambda^3\left[\sum_{x=3}^{\infty}\frac{\lambda^{x-3}}{(x-3)\,!}\right]$$

$$+ 7e^{-\lambda}\lambda^2\left[\sum_{x=2}^{\infty}\left(\frac{\lambda^{x-2}}{(x-2)\,!}\right)\right] + \lambda$$

$$= \lambda^4(e^{-\lambda}e^{+\lambda}) + 6\lambda^3(e^{-\lambda}e^{+\lambda}) + 7\lambda^2(e^{-\lambda}e^{+\lambda}) + \lambda = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$$

The four central moments are now obtained as follows :

$$\mu_2 = \mu_2' - \mu_1'^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda$$

Thus the mean and the variance of the Poisson distribution are each equal to λ.

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = (\lambda^3 + 3\lambda^2 + \lambda) - 3\lambda(\lambda^2 + \lambda) + 2\lambda^3 = \lambda.$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= (\lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda) - 4\lambda(\lambda^3 + 3\lambda^2 + \lambda) + 6\lambda^2(\lambda^2 + \lambda) - 3\lambda^4 = 3\lambda^2 + \lambda$$

Co-efficients of skewness and kurtosis are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda} \text{ and } \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{\lambda}}$$

Also $$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1}{\lambda} \text{ and } \gamma_2 = \beta_2 - 3 = \frac{1}{\lambda} \qquad \ldots(7\cdot 15)$$

Hence the Poisson distribution is always a skewed distribution.

Proceeding to the limit as $\lambda \to \infty$, we get

$$\beta_1 = 0 \text{ and } \beta_2 = 3$$

7·3·3. Mode of the Poisson Distribution

$$\frac{p(x)}{p(x-1)} = \frac{\dfrac{e^{-\lambda}\lambda^x}{x\,!}}{\dfrac{e^{-\lambda}\lambda^{x-1}}{(x-1)\,!}} = \frac{\lambda}{x} \qquad \ldots(7\cdot 16)$$

We discuss the following cases :

Case I. *When λ is not an integer.*

Let us suppose that S is the intergral part of λ.

$$\frac{p(1)}{p(0)} > 1, \quad \ldots, \frac{p(S-1)}{p(S-2)} > 1, \frac{p(S)}{p(S-1)} > 1,$$

and $$\frac{p(S+1)}{p(S)} < 1, \frac{p(S+2)}{p(S+1)} < 1, \ldots$$

Combining the above expressions into a single expression, we get

$p(0) < p(1) < p(2) \ldots < p(S-2) < p(S-1) < p(S) > p(S+1) > p(S+2) > \ldots$, which shows that $p(S)$ is the maximum value. Hence in this case the distribution is unimodal and the integral part of λ is the unique modal value.

Case II. *When $\lambda = k$ (say) is an integer.* Here we have

$$\frac{p(1)}{p(0)} > 1, \frac{p(2)}{p(1)} > 1, \ldots, \frac{p(k-1)}{p(k-2)} > 1$$

and $$\frac{p(k)}{p(k-1)} = 1, \frac{p(k+1)}{p(k)} < 1, \frac{p(k+2)}{p(k+1)} < 1, \ldots$$

$\therefore\ p(0) < p(1) < p(2) < \ldots < p(k-2) < p(k-1) = p(k) > p(k+1) > p(k+2) \ldots$

In this case we have two maximum values, *viz.*, $p(k-1)$ and $p(k)$ and thus the distribution is bimodal and two modes are at $(k-1)$ and k, *i.e.*, at $(\lambda - 1)$ and λ, (since $k = \lambda$).

7·3·4. Recurrence Relation for the Moments of the Poisson Distribution. By def.,

$$\mu_r = E\{X - E(X)\}^r = \sum_{x=0}^{\infty} (x-\lambda)^r p(x, \lambda)$$

$$= \sum_{x=0}^{\infty} (x-\lambda)^r \frac{e^{-\lambda}\lambda^x}{x!}$$

Differentiating with respect to λ, we get

$$\frac{d\mu_r}{d\lambda} = \sum_{x=0}^{\infty} r(x-\lambda)^{r-1}(-1)\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} \frac{(x-\lambda)^r}{x!}\left\{x\lambda^{x-1}e^{-\lambda} - \lambda^x e^{-\lambda}\right\}$$

$$= -r\sum_{x=0}^{\infty} (x-\lambda)^{r-1} \cdot \frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} \frac{(x-\lambda)^r}{x!}\left\{\lambda^{x-1}e^{-\lambda}(x-\lambda)\right\}$$

$$= -r\sum_{x=0}^{\infty} (x-\lambda)^{r-1}\frac{e^{-\lambda}\lambda^x}{x!} + \frac{1}{\lambda}\sum_{x=0}^{\infty} (x-\lambda)^{r+1} \cdot \frac{e^{-\lambda}\lambda^x}{x!}$$

$$\therefore \quad \frac{d\mu_r}{d\lambda} = -r\mu_{r-1} + \frac{1}{\lambda}\mu_{r+1}$$

$$\Rightarrow \quad \mu_{r+1} = r\lambda\mu_{r-1} + \lambda\frac{d\mu_r}{d\lambda} \qquad \ldots(7\cdot17)$$

Putting $r = 1, 2$ and 3 successively, we get

$$\mu_2 = r\mu_o + \lambda\frac{d\mu_1}{d\lambda} = \lambda \qquad (\because \mu_o = 1, \mu_1 = 0)$$

$$\mu_3 = 2\lambda\mu_1 + \lambda\frac{d\mu_2}{d\lambda} = \lambda,\ \mu_4 = 3\lambda\mu_2 + \lambda\frac{d\mu_3}{d\lambda} = 3\lambda^2 + \lambda$$

7·3·5. Moment Generating Function of the Poisson Distribution

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx}\cdot\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty}\frac{e^{-\lambda}(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda}\left\{1 + \lambda e^t + \frac{(\lambda e^t)^2}{2!} + \dots\right\} = e^{-\lambda}\cdot e^{\lambda e^t} = e^{\lambda(e^t-1)} \quad \dots(7\cdot18)$$

7·3·6. Characteristic Function of the Poisson Distribution

$$\phi_X(t) = \sum_{x=0}^{\infty} e^{itx}\cdot p(x) = \sum_{x=0}^{\infty} e^{itx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^{it})^x}{x!}$$

$$= e^{-\lambda}\, e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)} \quad \dots(7\cdot19)$$

7·3·7. Cumulants of the Poisson Distribution

$$K_X(t) = \log M_X(t) = \log\left[e^{\lambda(e^t-1)}\right] = \lambda(e^t - 1)$$

$$= \lambda\left[\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \dots + \frac{t^r}{r!} + \dots\right) - 1\right]$$

$$= \lambda\left[t + \frac{t^2}{2!} + \frac{t^3}{3!} + \dots + \frac{t^r}{r!} + \dots\right]$$

κ_r = rth cumulant = co-efficient of $\frac{t^r}{r!}$ in $K_X(t) = \lambda$

$$\Rightarrow \qquad \kappa_r = \lambda;\ r = 1, 2, 3, \dots \quad \dots(7\cdot19a)$$

Hence all the cumulants of the Poisson distribution are equal, each being equal to λ. In particular, we have

Mean $= \kappa_1 = \lambda$, $\mu_2 = \kappa_2 = \lambda$, $\mu_3 = \kappa_3 = \lambda$ and $\mu_4 = \kappa_4 + 3\kappa_2^2 = \lambda + 3\lambda^2$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\lambda + 3\lambda^2}{\lambda^2} = \frac{1}{\lambda} + 3$$

Remark. If m is the mean and σ is the s.d. of Poisson distribution with parameter λ, then

$$m\,\sigma\,\gamma_1\,\gamma_2 = \lambda\cdot\sqrt{\lambda}\cdot\sqrt{\beta_1}\,(\beta_2 - 3)$$

$$= \lambda\cdot\sqrt{\lambda}\cdot\frac{1}{\sqrt{\lambda}}\cdot\frac{1}{\lambda} = 1.$$

7·3·8. Additive or Reproductive Property of Independent Poisson Variates. *Sum of independent Poisson variates is also a Poisson variate. More elaborately, if X_i, $(i = 1, 2, \dots, n)$ are independent Poisson variates with param-*

ters λ_i; $i = 1, 2, ..., n$ *respectively, then* $\sum_{i=1}^{n} X_i$ *is also a Poisson variate with parameter* $\sum_{i=1}^{n} \lambda_i$.

Proof. $M_{X_i}(t) = e^{\lambda_i(e^t - 1)}$; $i = 1, 2, ..., n$

$$M_{X_1 + X_2 + \ldots + X_n}(t) = M_{X_1}(t)\, M_{X_2}(t) \ldots M_{X_n}(t),$$

[since X_i; $i = 1, 2, ..., n$ are independent]

$$= e^{\lambda_1(e^t - 1)}\, e^{\lambda_2(e^t - 1)} \ldots e^{\lambda_n(e^t - 1)}$$

$$= e^{(\lambda_1 + \lambda_2 + \ldots + \lambda_n)(e^t - 1)}$$

which is the m.g.f. of a Poisson variate with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_n$. Hence by uniqueness theorem of m.g.f.'s, $\sum_{i=1}^{n} X_i$ is also a Poisson variate with parameter $\sum_{i=1}^{n} \lambda_i$.

Remarks 1. In fact, the converse of the above result is also true *i.e., If* $X_1, X_2, ..., X_n$ *are independent and* $\sum_{i=1}^{n} X_i$ *has a Poisson distribution, then each of the random variables* $X_1, X_2, ..., X_n$ *has a Poisson distribution.*

Let X_1 and X_2 be independent *r.v.'s* so that $X_1 \sim P(\lambda_1)$ and $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$. Then we want to prove that $X_2 \sim P(\lambda_2)$.

Proof. Since X_1 and X_2 are independent, we have

$$M_{X_1 + X_2}(t) = M_{X_1}(t)\, M_{X_2}(t)$$

$$\Rightarrow \quad e^{(\lambda_1 + \lambda_2)(e^t - 1)} = e^{\lambda_1(e^t - 1)} \cdot M_{X_2}(t)$$

$$\Rightarrow \quad M_{X_2}(t) = e^{\lambda_2(e^t - 1)}$$

$\Rightarrow$ $X_2 \sim P(\lambda_2)$, by uniqueness theorem of m.g.f.

2. The difference of two independent Poisson variates is not a Poisson variate.

$$M_{X_1 - X_2}(t) = M_{X_1 + (-X_2)}(t) = M_{X_1}(t) \cdot M_{(-X_2)}(t),$$

(since X_1 and X_2 are independent).

$$\therefore \quad M_{X_1 - X_2}(t) = M_{X_1}(t)\, M_{X_2}(-t) \qquad [\because M_{cX}(t) = M_X(ct)]$$

$$= e^{\lambda_1(e^t - 1)} \cdot e^{\lambda_2(e^{-t} - 1)} = e^{\lambda_1(e^t - 1) + \lambda_2(e^{-t} - 1)}$$

which cannot be put in the form $e^{\lambda(e^t - 1)}$. Hence $(X_1 - X_2)$ is not a Poisson variate.

Moreover the difference $(X_1 - X_2)$ cannot be a Poisson variate is evident from the fact that it may have positive as well as negative values, while a Poisson variate is always non-negative.

7·3·9. Probability Generating Function of Poisson Distribution

$$\text{P.G.F. of } X = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!}\cdot s^k = \sum_{k=0}^{\infty} e^{-\lambda}\frac{(\lambda s)^k}{k!} = e^{-\lambda}e^{\lambda s} = e^{\lambda(s-1)} \qquad \text{...(7·20)}$$

Example 7·24. *A car hire firm has two cars which it fires out day by day. The number of demands for a car on each day is distributed as Poisson variate with mean 1·5., Calculate the proportion of days on which (i) neither car is used, and (ii) some demand is refused.* **[Meerut Univ. B.Sc. 1993]**

Solution. The proportion of days on which there are x demands for a car

$$= P\{\text{of } x \text{ demands in a day}\} = \frac{e^{-1\cdot 5}(1\cdot 5)^x}{x!},$$

since the number of demands for a car on any day is a Poisson variate with mean 1·5. Thus

$$P(X=x) = \frac{e^{-1\cdot 5}(1\cdot 5)^x}{x!};\quad x = 0, 1, 2, \ldots$$

(*i*) Proportion of days on which neither car is used is given by

$$P(X=0) = e^{-1\cdot 5} = \left[1 - 1\cdot 5 + \frac{(1\cdot 5)^2}{2!} - \frac{(1\cdot 5)^3}{3!} + \frac{(1\cdot 5)^4}{4!} - \ldots\right] = 0\cdot 2231$$

(*ii*) Proportion of days on which some demand is refused is

$$\begin{aligned}P(X>2) &= 1 - P(X \le 2)\\ &= 1 - [P(X=0) + P(X=1) + P(X=2)]\\ &= 1 - e^{-1\cdot 5}\left[1 + 1\cdot 5 + \frac{(1\cdot 5)^2}{2!}\right]\\ &= 1 - 0\cdot 2231 \times 3\cdot 625 = 0\cdot 19126\end{aligned}$$

Example 7·25. *A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality ?* **[Kanpur Univ. B.Sc. 1993]**

Solution. We are given $n = 100$.

Let p = Probability of a defective pin = 5% = 0·05

$\therefore$ λ = Mean number of defective pins in a box of 100

$= np = 100 \times 0\cdot 05 = 5$

Since 'p' is small, we may use Poisson distribution.

Probability of x defective pins in a box of 100 is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-5}5^x}{x!}; \; x = 0, 1, 2, \ldots$$

Probability that a box will fail to meet the guaranteed quality is

$$P(X > 10) = 1 - P(X \le 10) = 1 - \sum_{x=0}^{10} \frac{e^{-5}5^x}{x!} = 1 - e^{-5} \sum_{x=0}^{10} \frac{5^x}{x!}$$

Example 7·26. *Six coins are tossed 6,400 times. Using the Poisson distribution, find the approximate probability of getting six heads r times.*

Solution. The probability of obtaining six heads in one throw of six coins (a single trial), is $p = (½)^6$, assuming that head and tail are equally probable.

$$\therefore \qquad \lambda = np = 6400 \times (½)^6 = 100.$$

Hence, using Poisson probability law, the required probability of getting 6 heads r times is given by :

$$P(X = r) = \frac{e^{-\lambda} \cdot \lambda^r}{r!} = \frac{e^{-100} \cdot (100)^r}{r!}; \; r = 0, 1, 2, \ldots$$

Example 7·27. *In a book of 520 pages, 390 typo-graphical errors occur. Assuming Poisson law for the number of errors per page, find the probability that a random sample of 5 pages will contain no error.*

[Patna Univ. B.Sc. (Hons.), 1988]

Solution. The average number of typographical errors per page in the book is given by $\lambda = (390/520) = 0{\cdot}75$

Hence using Poisson probability law, the probability of x errors per page is given by : $P(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!} = \dfrac{e^{-0\cdot 75}(0{\cdot}75)^x}{x!}; x = 0, 1, 2, \ldots$

The required probability that a random sample of 5 pages will contain no error is given by : $[P(X = 0)]^5 = \left(e^{-0\cdot 75}\right)^5 = e^{-3\cdot 75}$

Example 7·28. *Suppose that the number of telephone calls coming into a telephone exchange between 10 A.M. and 11 A.M. say, X_1 is a random variable with Poisson distribution with parameter 2. Similarly the number of calls arriving between 11 A.M. and 12 noon say, X_2 has a Poisson distribution with parameter 6. If X_1 and X_2 are independent, what is the probability that more than 5 calls come in between 10 A.M. and 12 noon ?* **[Calicut U. B. Sc. Oct. 1992]**

Solution. Let $X = X_1 + X_2$. By the additive property of Poisson distribution, X is also a Poisson variate with parameter (say) $\lambda = 2 + 6 = 8$

Hence the probability of x calls in-between 10. A.M. and 12 noon is given by $P(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!} = \dfrac{e^{-8}8^x}{x!}; \; x = 0, 1, 2, \ldots$

Probability that more than 5 calls come in between 10 A.M. and 12 noon is given by

$$P(X > 5) = 1 - P(X \le 5) = 1 - \sum_{x=0}^{5} \frac{e^{-8}\, 8^x}{x\,!}$$
$$= 1 - 0\cdot 1912 = 0\cdot 8088$$

Example 7·29. *A Poisson distribution has a double mode at $x = 1$ and $x = 2$. What is the probability that x will have one or the other of these two values?*

Solution. We have proved that if the Poisson distribution is bimodal, then the two modes are at the points $x = \lambda - 1$ and $x = \lambda$. Since we are given that the two modes are at the points $x = 1$ and $x = 2$, we find that $\lambda = 2$.

$$\therefore \qquad P(X = x) = \frac{e^{-\lambda}\lambda^x}{x\,!} = \frac{e^{-2}\, 2^x}{x\,!}\,;\ x = 0, 1, 2, \dots$$

$$\Rightarrow \qquad P(X = 1) = e^{-2}\, 2$$

and
$$P(X = 2) = \frac{e^{-2} \cdot 2^2}{2\,!} = e^{-2} \cdot 2$$

Required probability $= P(X = 1) + P(X = 2) = 2e^{-2} + 2e^{-2} = 0{\cdot}542$

Example 7·30. *If X is a Poisson variate such that*

$$P(X = 2) = 9\, P(X = 4) + 90\, P(X = 6) \qquad \dots(*)$$

Find (i) λ, the mean of X, (ii) β, the coefficient of skewness.

[Delhi Univ. B. Sc. (Maths. Hons.) 1992, '87]

Solution. If X is a Poisson variate with parameter λ, then

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x\,!},\ x = 0, 1, 2, \dots;\ \lambda > 0$$

Hence (*) gives

$$\frac{e^{-\lambda} \cdot \lambda^2}{2\,!} = e^{-\lambda}\left[9\,\frac{\lambda^4}{4\,!} + 90\,\frac{\lambda^6}{6\,!}\right]$$
$$= \frac{e^{-\lambda}\lambda^2}{8}\,[3\lambda^2 + \lambda^4]$$

$$\Rightarrow \qquad \lambda^4 + 3\lambda^2 - 4 = 0$$

Solving as a quadratic in λ^2, we get

$$\lambda^2 = \frac{-3 \pm \sqrt{9 + 16}}{2} = \frac{-3 \pm 5}{2}$$

Since $\lambda > 0$, we get $\lambda^2 = 1 \quad \Rightarrow \quad \lambda = 1$

Hence mean $= \lambda = 1$, and $\mu_2 =$ Variance $= \lambda = 1$

Also $\quad \beta_1 =$ Coefficient of skewness $= \dfrac{1}{\lambda} = 1.$

Example 7·31. *If X and Y are independent Poisson variates such that*

$$P(X = 1) = P(X = 2)$$

and
$$P(Y = 2) = P(Y = 3) \qquad \dots(*)$$

Find the varaince of $X - 2Y$.

Solution. Let $X \sim P(\lambda)$ and $Y \sim P(\mu)$.

Then we have

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}, \; x = 0, 1, 2, \ldots ; \lambda > 0$$

and $$P(Y = y) = \frac{e^{-\mu} \cdot \mu^y}{y!}, \; y = 0, 1, 2, \ldots ; \mu > 0$$

Using (*), we get

$$\left.\begin{aligned} \lambda e^{-\lambda} &= \frac{\lambda^2 e^{-\lambda}}{2!} \\ \text{and} \quad \frac{\mu^2 e^{-\mu}}{2} &= \frac{\mu^3 e^{-\mu}}{3!} \end{aligned}\right\} \quad \ldots(**)$$

Solving (**), we get

$$\lambda e^{-\lambda} [\lambda - 2] = 0 \text{ and } \mu^2 e^{-\mu} [\mu - 3] = 0$$

$\Rightarrow$ $\lambda = 2$ and $\mu = 3$, since $\lambda > 0$, $\mu > 0$.

Now $\quad \text{Var}(X) = \lambda = 2$, and $\text{Var}(Y) = \mu = 3 \quad \ldots(***)$

$\therefore \quad \text{Var}(X - 2Y) = 1^2 \text{Var}(X) + (-2)^2 \cdot \text{Var}\, Y,$

covariance term vanishes since X and Y are independent.

Hence, on using (***), we get

$$\text{Var}(X - 2Y) = 2 + 4 \times 3 = 14$$

Example 7·32. *If X and Y are independent Poisson variates with means λ_1 and λ_2 respectively, find the probability that*

(*i*) $X + Y = k$, (*ii*) $X = Y$ [Delhi Univ. B. Sc. (Stat. Hons.), 1991]

Solution. We have

$$P(X = x) = \frac{e^{-\lambda_1} \cdot \lambda_1^x}{x!}, \; x = 0, 1, 2, 3, \ldots ; \lambda_1 > 0$$

and $$P(Y = y) = \frac{e^{-\lambda_2} \cdot \lambda_2^y}{y!}, \; y = 0, 1, 2, 3 \ldots ; \lambda_2 > 0$$

(*i*) $$P(X + Y = k) = \sum_{r=0}^{k} P(X = r \cap Y = k - r)$$

$$= \sum_{r=0}^{k} P(X = r)\, P(Y = k - r)$$

[$\because$ X and Y are independent]

$$= \sum_{r=0}^{k} \frac{e^{-\lambda_1} \lambda_1^r}{r!} \cdot \frac{e^{-\lambda_2} \cdot \lambda_2^{k-r}}{(k-r)!}$$

$$= e^{-(\lambda_1 + \lambda_2)} \sum_{r=0}^{k} \frac{\lambda_1^r \cdot \lambda_2^{k-r}}{r!\,(k-r)!}$$

$$= e^{-(\lambda_1+\lambda_2)}\left[\frac{\lambda_2^k}{k!}+\frac{\lambda_1\cdot\lambda_2^{k-1}}{1!(k-1)!}+\frac{\lambda_1^2\cdot\lambda_2^{k-2}}{2!(k-2)!}+\ldots+\frac{\lambda_1^k}{k!}\right]$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\left[\lambda_2^k+{}^kC_1\,\lambda_2^{k-1}\cdot\lambda_1+{}^kC_2\cdot\lambda_2^{k-2}\cdot\lambda_1^2+\ldots+\lambda_1^k\right]$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\times(\lambda_1+\lambda_2)^k,\ k=0,1,2,\ldots$$

which is the probability function of Poisson distribution with parameter $\lambda_1+\lambda_2$.

Aliter. Since $X \sim P(\lambda_1)$ and $Y \sim P(\lambda_2)$ are independent, by the additive property of Poisson distribution $X+Y \sim P(\lambda_1+\lambda_2)$. Hence

$$P(X+Y=k)=\frac{e^{-(\lambda_1+\lambda_2)}\times(\lambda_1+\lambda_2)^k}{k!};\ k=0,1,2,\ldots$$

(*ii*) $$P(X=Y)=\sum_{r=0}^{\infty}P(X=r\cap Y=r)$$

$$=\sum_{r=0}^{\infty}P(X=r)\,P(Y=r)$$

[$\because X$ and Y are independent]

$$=e^{-(\lambda_1+\lambda_2)}\sum_{r=0}^{\infty}\frac{(\lambda_1\lambda_2)^r}{(r!)^2}$$

Example 7·33. *Show that in a Poisson distribution with unit mean, mean deviation about mean is $(2/e)$ times the standard deviation.*

[Patna Univ. B. Sc. (Stat. Hons.) 1992; Delhi Univ. B.Sc. (Stat. Hons.), 1993]

Solution. Here we are given $\lambda=1$.

$$\therefore\quad P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-1}\cdot 1}{x!}=\frac{e^{-1}}{x!};\ x=0,1,2,\ldots$$

Mean deviation about mean 1 is

$$E(|X-1|)=\sum_{x=0}^{\infty}|x-1|\,p(x)=e^{-1}\sum_{x=0}^{\infty}\frac{|x-1|}{x!}$$

$$=e^{-1}\left[1+\frac{1}{2!}+\frac{2}{3!}+\frac{3}{4!}+\ldots\right]$$

We have $$\frac{n}{(n+1)!}=\frac{(n+1)-1}{(n+1)!}=\frac{1}{n!}-\frac{1}{(n+1)!}$$

$\therefore$ Mean deviation about mean

$$= e^{-1}\left[1 + \left(1 - \frac{1}{2!}\right) + \left(\frac{1}{2!} - \frac{1}{3!}\right) + \left(\frac{1}{3!} - \frac{1}{4!}\right) + \cdots\right]$$

$$= e^{-1}(1 + 1) = \frac{2}{e} \times 1 = \frac{2}{e} \times \text{standard deviation},$$

since for the Poisson distribution, variance = mean = 1 (given).

Example 7·34. *Let $X_1, X_2, \ldots, X_n$ be identically and independently distributed Bin (1, p) variates. Let $S_n = \sum_{j=1}^{n} X_j$ and $M_n(t)$ be the m.g.f. of S_n. Find $\lim_{n \to \infty} M_n(t)$, using $np = \lambda$ (const.)* **[Delhi Univ. B. Sc. (Maths Hons.), 1989]**

Solution. Since X_i, $i = 1, 2, \ldots n$ are i.i.d. binomial variates $B(1, p)$,

$S_n = \sum_{j=1}^{n} X_j$, is a binomial $B(n, p)$ variate.

$$\therefore M_n(t) = M.g.f. \text{ of } S_n = (q + pe^t)^n = \left[1 + (e^t - 1)p\right]^n$$

If we take $np = \lambda \Rightarrow p = \lambda/n$ and let $n \to \infty$, we get

$$\lim_{n \to \infty} M_n(t) = \lim_{n \to \infty}\left[1 + \frac{(e^t - 1)\lambda}{n}\right]^n = \exp\left[\lambda(e^t - 1)\right],$$

which is the m.g.f. of Poisson distribution with parameter λ. Hence by uniqueness theorem of m.g.f., $S_n = \sum_{j=1}^{n} X_j \to P(\lambda)$, as $n \to \infty$, with $np = \lambda$ (fixed).

Example. 7·35. *(a) If X is a Poisson variate with mean m, show that the expectation of e^{-kX} is $\exp\left[-m(1 - e^{-k})\right]$.* **[Nagpur Univ. B.Sc. 1993]**

Hence show that, if $\bar{X}$ is the arithmetic mean of n independent random variables $X_1, X_2, \ldots, X_n$, each having Poisson distribution with parameter m, then $e^{-\bar{x}}$ as an estimate of e^{-m} is biased, although $\bar{X}$ is an unbaised estimate of m.

(b) If X is a Poisson variate with mean m, what would be the expectation of $e^{-kx}\, kX$, k being a constant.

Solution.

$$E(e^{-kX}) = \sum_{x=0}^{\infty} e^{-kx} p(x) = \sum_{x=0}^{\infty} e^{-kx} \cdot \frac{e^{-m} m^x}{x!} = e^{-m} \sum_{x=0}^{\infty} \frac{(me^{-k})^x}{x!}$$

$$= e^{-m}\left[1 + me^{-k} + \frac{(me^{-k})^2}{2!} + \cdots\right]$$

$$= e^{-m} e^{me^{-k}} = e^{-m(1-e^{-k})} \qquad ...(*)$$

We have

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

Since X_i; $i = 1, 2, ..., n$ is a Poisson variate with parameter m, $E(X_i) = m$.

$$\therefore\ E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} m = \frac{1}{n}\, nm = m$$

Hence $\bar{X}$ is an unbaised estimate of m.

Now $\quad E(e^{-\bar{X}}) = E\left[\exp\left(\frac{-1}{n}\sum_{i=1}^{n} X_i\right)\right]$

$$= E\left(e^{-X_1/n} \cdot e^{-X_2/n} \cdots e^{-X_n/n}\right),$$

$$= E(e^{-X_1/n})\, E(e^{-X_2/n}) \ldots E(e^{-X_n/n}),$$

(since $X_1, X_2, ..., X_n$ are independent)

$$\therefore \quad E(e^{-\bar{X}}) = \prod_{i=1}^{n} E(e^{-X_i/n}) \qquad ...(**)$$

Using (*) with $k = 1/n$, we get

$E(e^{-X_i/n}) = e^{-m(1-e^{-1/n})}$, (since X_i is a Poisson variate with parameter m)

$$\therefore\ E(e^{-\bar{X}}) = \prod_{i=1}^{n}\left[\exp\left\{-m(1-e^{-1/n})\right\}\right] = \left[\exp\left\{-m(1-e^{-1/n})\right\}\right]^{n}$$

$$= \exp\left\{-mn(1-e^{-1/n})\right\} \neq e^{-m}$$

Hence e^{-X} is not an unbaised estimated of e^{-m}, though $\bar{X}$ is an unbiased estimate of m.

$$(b)\ E(e^{-kX} kX) = \sum_{x=0}^{\infty} e^{-kx} kx \cdot p(x) = k \sum_{x=1}^{\infty} e^{-kx}\, x\, \frac{e^{-m} m^x}{x!}$$

$$= ke^{-m} \sum_{x=1}^{\infty} \frac{(me^{-k})^x}{(x-1)!} = ke^{-m} me^{-k} \sum_{x=1}^{\infty} \frac{(me^{-k})^{x-1}}{(x-1)!}$$

$$= mke^{-m-k}\left\{1 + me^{-k} + \frac{(me^{-k})^2}{2!} + \ldots\right\}$$

$$= mk e^{-m-k} \cdot e^{me^{-k}} = mk \exp\left[\left\{m(e^{-k}-1)\right\} - k\right]$$

Example 7·36. *If X and Y are independent Poisson variates with means* m_1 and m_2 *respectively, prove that the probability that* $X - Y$ *has the value 'r' is the co-efficient of* t^r *in*

$$\exp\left\{m_1 t + m_2 t^{-1} - m_1 - m_2\right\}$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1991, '89]

Solution. Since X and Y are independent Poisson variates with means m_1 and m_2 respectively,

$$\left.\begin{array}{ll} & P(X = x) = \dfrac{e^{-m_1} m_1^x}{x!}\,;\ x = 0, 1, 2, \ldots \infty \\ \text{and} & P(Y = y) = \dfrac{e^{-m_2} m_2^y}{y!}\,;\ y = 0, 1, 2, \ldots \infty \end{array}\right\} \quad \ldots(1)$$

$$P(X - Y = r) = \sum_{s=0}^{\infty} P(X = r + s \cap Y = s) = \sum_{s=0}^{\infty} P(X = r + s)\, P(Y = s)$$

$$= \sum_{s=0}^{\infty} \frac{e^{-m_1} \cdot m_1^{r+s}}{(r+s)!} \cdot \frac{e^{-m_2} m_2^s}{s!} \quad \ldots[\text{From (1)}]$$

$$= e^{-m_1 - m_2} \sum_{s=0}^{\infty} \frac{m_1^{r+s} m_2^s}{(r+s)!\, s!} \quad \ldots(2)$$

We have $e^{m_1 t + m_2 t^{-1}} = e^{m_1 t} \times e^{m_2 t^{-1}}$

$$= \left\{1 + m_1 t + \frac{(m_1 t)^2}{2!} + \ldots + \frac{(m_1 t)^{r+s}}{(r+s)!} + \ldots\right\}$$

$$\times \left\{1 + m_2 t^{-1} + \frac{(m_2 t^{-1})^2}{2!} + \ldots + \frac{(m_2 t^{-1})^s}{s!} + \ldots\right\}$$

$\therefore$ Co-efficient of t^r in $e^{m_1 t + m_2 t^{-1}} = \displaystyle\sum_{s=0}^{\infty} \frac{m_1^{r+s} m_2^s}{(r+s)!\, s!}$

Hence from (2), we get

$$P(X - Y = r) = e^{-m_1 - m_2} \times \text{Coefficient of } t^r \text{ in } e^{m_1 t + m_2 t^{-1}},$$

$$= \text{Coefficient of } t^r \text{ in } e^{-m_1 - m_2 + m_1 t + m_2 t^{-1}}$$

which is the required result.

Example 7·37. *If X is a Poisson variate with mean m, show that $\dfrac{X - m}{\sqrt{m}}$ is a variable with mean zero and variance unity. Find the M.G.F. for this variable and show that it approaches $e^{t^2/2}$ as $m \to \infty$. Also interpret the result.*

[Delhi Univ. B. Sc. (Stat. Hons.), 1987]

Solution. Let $Y = \dfrac{X - m}{\sqrt{m}}$

$$\therefore \quad E(Y) = E\left(\frac{X - m}{\sqrt{m}}\right) = \frac{1}{\sqrt{m}} E(X - m) = 0$$

$$V(Y) = E\left(\frac{X-m}{\sqrt{m}}\right)^2 = \frac{1}{m}E(X-m)^2 = \frac{1}{m}\mu_2 = 1$$

$$\begin{aligned}
M.G.F. \text{ of } Y = M_Y(t) &= E(e^{tY}) = E\left[e^{t(X-m)/\sqrt{m}}\right] \\
&= e^{-t\sqrt{m}}\left[E(e^{tX/\sqrt{m}})\right] \\
&= e^{-t\sqrt{m}}\sum_{x=0}^{\infty}\frac{e^{-m}m^x}{x!}\cdot e^{tx/\sqrt{m}} \\
&= e^{-t\sqrt{m}}\cdot e^{-m}\sum_{x=0}^{\infty}\frac{(me^{t/\sqrt{m}})^x}{x!} \\
&= e^{-m-t\sqrt{m}}\left[1+\frac{me^{t/\sqrt{m}}}{1!}+\frac{(me^{t/\sqrt{m}})^2}{2!}+\dots\right] \\
&= e^{-m-t\sqrt{m}}\cdot\exp(m e^{t/\sqrt{m}}) = \exp\left[-m-t\sqrt{m}+me^{t/\sqrt{m}}\right] \\
&= \exp\left[-m-t\sqrt{m}+m\left(1+\frac{t}{\sqrt{m}}+\frac{t^2}{2!\,m}+\frac{t^3}{3!\,m^{3/2}}+\dots\right)\right] \\
&= \exp\left[\frac{1}{2}t^2+\frac{1}{3!}\frac{t^3}{\sqrt{m}}+\dots\right]
\end{aligned}$$

Now proceeding to limit as $m \to \infty$, we get

$$\lim_{m\to\infty} M_Y(t) = e^{t^2/2} \qquad \dots(*)$$

Interpretation. (*) is the m.g.f. of Standard Normal Variate [c.f. Remark to § 8·2·5]. Hence by uniqueness theorem of m.g.f.'s, standard Poisson variate tends to standard normal variate as $m \to \infty$. Hence Poisson distribution tends to Normal distribution for large values of parameter m.

Example 7·38. *Deduce the first four moments about the mean of the Poisson distribution from those of the Binomial distribution.*

Solution. The first four central moments of the binomial distribution are

$$\left[\begin{array}{ll} \mu_1 = 0, & \text{Mean} = np \\ \mu_2 = npq, & \mu_3 = npq(q-p) \text{ and} \\ \mu_4 = npq(1-6pq) + 3n^2p^2q^2 & \end{array}\right] \qquad \dots(*)$$

Poisson distribution is a limiting form of the binomial distribution under the following conditions :

(*i*) $n \to \infty$, (*ii*) $p \to 0$, *i.e.*, $q \to 1$, and (*iii*) $np = \lambda$, (say), is finite.

Using these conditions, we get from (*) the moments of the Poisson distribution as

$$\begin{aligned}
\mu_1 &= 0 \\
\text{Mean} &= \lim(np) = \lambda \\
\mu_2 &= \lim(npq) = \lim(np)\cdot\lim(q) = \lambda\cdot 1 = \lambda \\
\mu_3 &= \lim[npq(q-p)] = \lambda\cdot 1(1-0) = \lambda \\
\mu_4 &= \lim[npq(1-6pq) + 3(np)^2q^2]
\end{aligned}$$

$$= [\lambda \cdot 1\ (1 - 6 \cdot 0 \cdot 1) + 3\lambda^2 \cdot 1] = \lambda + 3\lambda^2$$

Example 7·39. *If X is a Poisson variate with parameter m and Y is another discrete variable whose conditional distribution for a given X is given by*

$$P(Y = r \mid X = x) = \binom{x}{r} p^r (1 - p)^{x-r}\ ;\ 0 < p < 1,\ r = 0, 1, 2, \ldots, x$$

then show that the unconditional distribution of Y is a Poisson distribution with parameter mp.

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, Shivaji U.B.Sc. Nov. 1992]

Solution. We are given that

$$P(X = x) = \frac{e^{-m} m^x}{x!}\ ;\ x = 0, 1, 2, \ldots$$

and
$$P(Y = r \mid X = x) = \binom{x}{r} p^r (1 - p)^{x-r}\ ;\ r \le x$$

$$\therefore\quad P(X = x \cap Y = r) = P(X = x)\, P(Y = r \mid X = x)$$
$$= \frac{e^{-m} m^x}{x!} \binom{x}{r} p^r (1 - p)^{x-r}$$

$\therefore\ P(Y = r) =$ The unconditional distribution of Y.

$$= \sum_{x=r}^{\infty} \left[\frac{e^{-m} m^x}{x!} \cdot \binom{x}{r} p^r (1 - p)^{x-r} \right]$$

$$= e^{-m} \left[\sum_{x=r}^{\infty} \binom{x}{r} \frac{p^r m^x (1 - p)^{x-r}}{x!} \right]$$

$$= e^{-m} \left[\sum_{x=r}^{\infty} \frac{m^x}{x!} \cdot \frac{x!}{r!\,(x - r)!} p^r (1 - p)^{x-r} \right]$$

$$= \frac{e^{-m}}{r!} \left[\sum_{x=r}^{\infty} \frac{m^x}{(x - r)!} p^r (1 - p)^{x-r} \right]$$

$$= \frac{e^{-m} (mp)^r}{r!} \left[\sum_{x=r}^{\infty} \frac{m^{x-r} (1 - p)^{x-r}}{(x - r)!} \right]$$

$$= \frac{e^{-m} (mp)^r}{r!} \left[\sum_{x=r}^{\infty} \frac{\{m(1 - p)\}^{x-r}}{(x - r)!} \right]$$

$$= \frac{e^{-m} (mp)^r}{r!} e^{m(1-p)} = \frac{e^{-mp} (mp)^r}{r!}\ ;\ r = 0, 1, 2, \ldots$$

Hence Y is a Poisson variate with parameter mp.

Example 7·40. *If X and Y are independent Poisson variates, show that the conditional distribution of X given X + Y, is binomial.*

[Madras Univ. B.Sc. Main 1992; Delhi Univ. B. Sc. (Maths Hons.), 1988]

Solution. Let X and Y be independent Poisson variates with parameters λ and μ respectively. Then $X+Y$ is also a Poisson variate with parameter $\lambda+\mu$.

$$P[X=r \mid (X+Y=n)] = \frac{P(X=r \cap X+Y=n)}{P(X+Y=n)} = \frac{P(X=r \cap Y=n-r)}{P(X+Y=n)}$$

$$= \frac{P(X=r)P(Y=n-r)}{P(X+Y=n)} \quad \text{[since } X \text{ and } Y \text{ are indepdent]}$$

$$\therefore P[X=r \mid (X+Y=n)] = \frac{e^{-\lambda}\dfrac{\lambda^r}{r!}\cdot e^{-\mu}\dfrac{\mu^{n-r}}{(n-r)!}}{\dfrac{e^{-(\lambda+\mu)}(\lambda+\mu)^n}{n!}}$$

$$= \frac{n!}{r!(n-r)!}\left(\frac{\lambda}{\lambda+\mu}\right)^r\left(\frac{\mu}{\lambda+\mu}\right)^{n-r}$$

$$= \binom{n}{r}p^r q^{n-r}, \text{ where } p=\frac{\lambda}{\lambda+\mu}, q=1-p$$

Hence the conditional distribution of X given $X+Y=n$, is a binomial distribution with parameters n and $p=\lambda/(\lambda+\mu)$.

Example 7·41. *If X is a Poisson variate with parameter m and μ_r is the rth central moment, prove that*

$$m\,[{}^rC_1\,\mu_{r-1} + {}^rC_2\,\mu_{r-2} + \ldots + {}^rC_r\,\mu_0] = \mu_{r+1}.$$

[Delhi Univ. B.Sc. (Stat. Hons.) 1990]

Solution Since $X \sim P(m)$, its probability function is given by

$$p(x) = \frac{e^{-m}\cdot m^x}{x!}, \; x=0,1,2,\ldots; m>0$$

By definition,

$$\mu_{r+1} = E[X-E(X)]^{r+1} = E[X-m]^{r+1}$$

$$= \sum_{x=0}^{\infty}(x-m)^{r+1}p(x)$$

$$= \sum_{x=0}^{\infty}(x-m)^r(x-m)\frac{e^{-m}\cdot m^x}{x!}$$

$$= \sum_{x=0}^{\infty}\frac{x(x-m)^r e^{-m}m^x}{x!} - m\sum_{x=0}^{\infty}(x-m)^r\cdot\frac{e^{-m}m^x}{x!}$$

$$= \sum_{x=1}^{\infty} \frac{(x-m)^r e^{-m} m^x}{(x-1)!} - m\mu_r$$

$$= \sum_{y=0}^{\infty} \frac{(y-m+1)^r \cdot e^{-m} \cdot m^{y+1}}{y!} - m\mu_r, \qquad (x-1=y)$$

$$= m \cdot \sum_{y=0}^{\infty} (y-m+1)^r \cdot p(y) - m\mu_r$$

$$= m \sum_{y=0}^{\infty} \Big[(y-m)^r + {}^rC_1 (y-m)^{r-1} + {}^rC_2 (y-m)^{r-2}$$

$$+ \ldots + {}^rC_{r-1}(y-m) + 1 \Big] p(y) - m\mu_r$$

$$= m[\mu_r + {}^rC_1 \mu_{r-1} + {}^rC_2 \mu_{r-2} + \ldots + {}^rC_r \mu_0] - m\mu_r$$

$$= m[{}^rC_1 \mu_{r-1} + {}^rC_2 \mu_{r-2} + \ldots + {}^rC_r \mu_0].$$

Example 7·42. *If X has a Poisson distribution with parameter λ, show that the distribution function of X is given by*

$$F(x) = \frac{1}{\Gamma(x+1)} \int_{\lambda}^{\infty} e^{-t} t^x \, dt \; ; \; x = 0, 1, 2, \ldots$$

[Delhi Univ. M. Sc. (Stat) 1986]

Solution. If X is a Poisson variate, then

$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!} \; ; \; x = 0, 1, 2, \ldots \qquad (*)$$

Consider the incomplete gamma integral;

$$I_x = \frac{1}{x!} \int_{\lambda}^{\infty} e^{-t} t^x \, dt \; ; \quad (x \text{ is a positive integer})$$

$$= \left| -\frac{e^{-t} t^x}{x!} \right|_{\lambda}^{\infty} + \frac{1}{(x-1)!} \int_{\lambda}^{\infty} e^{-t} t^{x-1} \, dt$$

$$= \frac{e^{-\lambda} \lambda^x}{x!} + I_{x-1} \qquad (**)$$

which is a reduction formula for I_x.

Repeated applications of (**) gives

$$I_x = \frac{e^{-\lambda} \lambda^x}{x!} + \frac{e^{-\lambda} \lambda^{x-1}}{(x-1)!} + \ldots + \frac{e^{-\lambda} \lambda}{1!} + I_0$$

$$\text{But } I_0 = \int_{\lambda}^{\infty} e^{-t} \, dt = \left| -e^{-t} \right|_{\lambda}^{\infty} = e^{-\lambda}$$

$$\therefore \quad I_x = e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^x}{x!} e^{-\lambda}$$

$$= P(X=0) + P(X=1) + \ldots + P(X=x) \qquad \text{[From (*)]}$$

$$= P(X \leq x) = F(x)$$

where $F(\cdot)$ is the distribution function of the $r.v. X$.

$$\Rightarrow \qquad F(x) = \frac{1}{x!}\int_{\lambda}^{\infty} e^{-t} t^x \, dt = \frac{1}{\Gamma(x+1)}\int_{\lambda}^{\infty} e^{-t} t^x \, dt$$

($\because \Gamma(x+1) = x!$, since x is a positive integer.)

Remark. This result is of great practical utility. It enables us to represent the cumulative Poisson probabilities (which are generally tedious to compute numerically) in terms of incomplete gamma integral, the values of which are tabulated for different values of λ by Karl Pearson in his Tables of Incomplete Γ-Functions.

7·3·10. Recurrence Formula for the Probabilities of Poisson Distribution. *(Fitting of Poisson Distribution).* For a Poisson distribution with parameter λ, we have

$$p(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}\ ; \ x = 0, 1, 2, \ldots, \infty$$

and $$P(x+1) = \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}\ ; \ x = 0, 1, 2, \ldots, \infty$$

$$\therefore \qquad \frac{p(x+1)}{p(x)} = \frac{\lambda}{(x+1)} \Rightarrow p(x+1) = \frac{\lambda}{x+1}p(x) \qquad \ldots(17\cdot20)$$

which is the required recurrence formula.

This formula provides us a very convenient method of graduating the given data by a Poisson distribution. The only probability we need to calculate is $p(0)$ which is given by $p(0) = e^{-\lambda}$, where λ is estimated from the given data. The other probabilities, *viz.*, $p(1), p(2)$.... can now be easily obtained as explained below:

$$p(1) = [p(x+1)]_{x=0} = \left[\frac{\lambda}{x+1}\right]_{x=0} p(0),$$

$$p(2) = [p(x+1)]_{x=1} = \left[\frac{\lambda}{x+1}\right]_{x=1} p(1),$$

$$p(3) = [p(x+1)]_{x=2} = \left[\frac{\lambda}{x+1}\right]_{x=2} p(2),$$

and so on.

Example 7·43. *After correcting 50 pages of the proof of a book, the proof reader finds that there are, on the average, 2 errors per 5 pages. How many pages would one expect to find with 0, 1, 2, 3 and 4 errors, in 1000 pages of the first print of the book ? (Given that $e^{-0\cdot4} = 0\cdot6703$)*

[Nagpur Univ. M.A. (Eco.), 1989]

Solution. Let the random variable X denote the number of errors per page. Then the mean number of errors per page is given by :

$$\lambda = 2/5 = 0{\cdot}4$$

Using Poisson probability law, probability of x errors per page is given by:

$$P(X = x) = p(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0\cdot4}(0{\cdot}4)^x}{x!};\ x = 0, 1, 2, \ldots$$

Expected number of pages with x errors per page in a book of 1000 pages are :

$$1000 \times P(X = x) = 1000 \times \frac{e^{-0\cdot4}(0{\cdot}4)^x}{x!};\ x = 0, 1, 2, \ldots$$

Using the recurrence formula (17·20), various probabilities can be easily calculated as shown in the following table.

No. of errors per page (X)	*Probability* $p(x)$	*Expected number of pages* $1000\,p(x)$
0	$p(0) = e^{-0\cdot4} = 0{\cdot}6703$	$670{\cdot}3 \simeq 670$
1	$p(1) = \frac{0\cdot4}{0+1}\,p(0) = 0{\cdot}26812$	$268{\cdot}12 \simeq 268$
2	$p(2) = \frac{0\cdot4}{1+1}\,p(1) = 0{\cdot}053624$	$53{\cdot}624 \simeq 54$
3	$p(3) = \frac{0\cdot4}{2+1}\,p(2) = 0{\cdot}0071298$	$7{\cdot}1298 \simeq 7$
4	$p(4) = \frac{0\cdot4}{3+1}\,p(3) = 0{\cdot}00071298$	$0{\cdot}71298 \simeq 1$

Example 7·44. *Fit a Poisson distribution to the following data which gives the number of doddens in a sample of clover seeds.*

No. of doddens: (x)	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Observed frequency: (f)	*56*	*156*	*132*	*92*	*37*	*22*	*4*	*0*	*1*

Solution.

$$\text{Mean} = \frac{1}{N}\sum fx = \frac{986}{500} = 1{\cdot}972$$

Taking the mean of the given distribution as the mean of the Poisson distribution we want to fit, we get $\lambda = 1{\cdot}972$,

and
$$p(x) = \frac{e^{-\lambda}\cdot\lambda^x}{x!};\ x = 0, 1, 2, \ldots, \infty$$

$$p(0) = e^{-\lambda} = e^{-1\cdot972}$$

$\therefore \log_{10} p(0) = -1{\cdot}972 \log_{10} e = -1{\cdot}972 \times 0{\cdot}43429$

$= -0{\cdot}856419 = \bar{1}{\cdot}143581$

$\therefore \quad p(0) = 0{\cdot}1392$

Using the recurrence formula (17·20) the various probabilities, *viz.*, $p(1), p(2), \ldots$, can be easily calculated as shown in the following table :

x	$\frac{\lambda}{x+1}$	$p(x)$	*Expected frequency* $N.p(x)$
0	1·972	0·13920	69·6000
1	0·986	0·27455	137·2512
2	0·657	0·27006	135·3296
3	0·493	0·17793	88·9566
4	0·394	0·10964	43·8556
5	0·328	0·03459	17·2966
6	0·281	0·01137	5·6846
7	0·247	0·00320	1·6013
8	0·219	0·00078	0·3942

Since frequencies are always integers, therefore by converting them to nearest integers, we get

Observed frequency :	56	156	132	92	37	22	4	0	1
Expected frequency :	70	137	135	89	44	17	6	2	0

Remark. In rounding the figures to the nearest integer it has to be kept in mind that the total of the observed and the expected frequencies should be same.

EXERCISE 7 (b)

1. (*a*) Derrive Poisson distribution as a limiting form of a binomial distribution. **[Madras Univ. B. E., Dec. 1991]**

Hence find β_1 and β_2 of the distribution.

Give some examples of the occurrence of Poisson distribution in different fields.

(*b*) State and prove the reproductive property of the Poisson distribution. Show that the mean and variance of the Poisson distribution are equal.

Find the mode of the Poisson distribution with mean value 5.

(*c*) Prove that under certain conditions to be stated by you, the number of telephone calls on a trunkline in a given interval of time has a Poisson distribution. **[Calcutta Univ. B.Sc. (Maths Hons.), 1989]**

(*d*) Show that for a Poisson distribution, the coefficient of variation is the reciprocal of the standard deviation.

2. (*a*) If two independent variables X_1 and X_2 have Poisson distribution with means λ_1 and λ_2 respectively, then show that their sum $X_1 + X_2$ is a Poisson variate with mean $\lambda_1 + \lambda_2$.

Does the difference of two independent Poisson variates follow a Poisson distribution ? Give reasons. **[Sri Venketeswara Univ. B.Sc., 1991]**

(*b*) Prove that the sum of two independent Poisson variates is a Poisson variate. Is the result true for the difference also ? Give reasons.

[Delhi Univ. B.Sc. (Stat. Hons.) 1989]

(*c*) If $X_1, X_2, ..., X_k$ are independent random variables following the Poisson law with parameter $m_1, m_2, ..., m_k$ respectively, show that $\sum_{i=1}^{k} X_i$ follows the Poisson law with parameter $\sum_{i=1}^{k} m_i$ **[Madras Univ. B. E., 1993]**

3. (*a*) Prove the recurrence relation between the moments of Poisson distribution

$$\mu_{r+1} = \lambda \left(r\mu_{r-1} + \frac{d\mu_r}{d\lambda} \right), \text{ where } \mu_r = \sum_{j=0}^{\infty} \frac{e^{-\lambda} \lambda^j}{j!} (j-\lambda)^r$$

where μ_r is the *r*th moment about the mean λ. Hence obtain the skewness and kurtosis of Poisson distribution.

[Delhi Univ. B. Sc. (Stat. Hons.) 1989,' 86; Utkal Univ. B. Sc. 1993]

(*b*) Let X have a Poisson distribution with parameter $\lambda > 0$. If r is a non-negative integer and if $\mu_r' = E(X^r)$, prove that

$$\mu'_{r+1} = \lambda \left(\mu'_r + \frac{d\mu'_r}{d\lambda} \right)$$

[Madras Univ. B. Sc. Nov. 1988]

4. What do you understand by (*i*) cumulants, (*ii*) cumulative function. Obtain the cumulative function of a Poisson distribution with parameter λ. Hence or otherwise show that for a Poisson distribution with parameter λ, all the cumulants are λ.

5. For the Poisson distribution with parameter λ, show that the *r*th factorial moment $\mu'_{(r)}$ is given by $\mu'_{(r)} = \lambda^r$

Show further that $\mu_{(2)} = \lambda$, $\mu_{(3)} = -2\lambda$ and $\mu_{(4)} = 3\lambda(\lambda + 2)$

6. (*a*) If X and Y are independent *r.v. s.*' so that $X \sim P(\lambda)$ and $X + Y \sim P(\lambda + \mu)$, find the distribution of Y. **[Ans.** $Y \sim P(\mu)$**]**

(*b*) If $X \sim P(\lambda)$, find

(*i*) Karl Pearson's coefficient of skewness

(*ii*) Moment measure of skewness.

Is Poisson distribution positively skewed or negatively skewed ?

7. (*a*) It is known that the probability that an item produced by a certain machine will be defective is 0·01. By applying Poisson's approximation, show that the probability that random sample of 100 items selected at random from the total output will contain no more than one defective item is $2/e$.

(*b*) The probability of success in a trail is known to be 10^{-4}. It is possible to repeat the trial independently any desired number of times. Do you think that the number of successes in a series of trials, if the number of trials in the series increases indefinitely, will tend to follow a Poisson distribution ? Give your reasons.

(*c*) The probability of getting no misprint in a page of a book is e^{-4}. What is the probability that a page contains more than 2 misprints ? [State the assumptions you make in solving this problem.] **[Bombay Univ. B.Sc., 1989]**

8. In a certain factory turning out optical lenses, there is a small chance 1/500 for any lens to be defective. The lenses are supplied in a packet of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective, two defective and three defective lenses in a consignment of 20,000 packets.

Ans. 19604, 392, 4 and 0 packets.

9. Red blood cell deficiency may be determined by examining a specimen of the blood under a microscope. Suppose a certain small fixed volume contains on the average 20 red cells for normal persons. Using Poisson distribution, obtain the probability that a specimen from a normal person will contain less than 15 red cells.

Ans. $\sum_{x=0}^{14} \{e^{-20} (20)^x/x\,!\}$

10. Assuming that the chance of a traffic accident in a day in a street of Delhi is 0·001, on how many days out of a trial of 1,000 days can we expect :

(*i*) no accident

(*ii*) more than three accidents, if there are 1,000 such streets in the whole city ?

11. Patients arrive randomly and independently at a doctor's surgery from 8·0 A.M. at an average rate of one in five minutes.The waiting room holds 2 persons. What is the probability that the room will be full when the doctor arrives at 9·0 A.M. (Estimate the probability to an accuracy of 5 per cent.)

Ans. 53·84 %

12. An office switchboard receives telephone calls at the rate of 3 calls per minute on an average. What is the probability of receiving (*i*) no calls in a one-minute interval, (*ii*) at the most 3 calls in a 5-minute interval ?

Ans. (*i*) 0·0323, (*ii*) 0

13. A hospital switchboard receives an average of 4 emergency calls in a 10-minute interval. What is the probability that (*i*) there are at the most 2 emergency calls in a 10-minute interval, (*ii*) there are exactly 3 emergency calls in a 10-minute interval ?

Ans. (*i*) 13^{-4}, (*ii*) $(32/3)\,e^{-4}$

14. (*a*) A distributor of bean seeds determines from extensive tests that 5% of large batch of seeds will not germinate. He sells the seeds in packets of 200 and guarantees 90% germination. Determine the probability that a particular packet will violate the guarantee.

Ans. $1 - \sum_{r=0}^{10} (e^{-10}\, 10^r / r!)$

(*b*) In an automatic telephone exchange the probability that any one call is wrongly connected is 0·001. What is the minimum number of independent calls required to ensure a probability of 0·90, that at least one call is wrongly connected?

15. (*a*) Fit a Poisson distribution to the following data with respect to the number of red blood corpuscles (x) per cell :

x :	0	1	2	3	4	5
Number of cells f :	142	156	69	27	5	1

(*b*) Data was collected over a period of 10 years, showing number of deaths from horse kicks in each of the 20 army corps. From the 200 corps-years, the distribution of deaths was as follows :

No. of deaths :	0	1	2	3	4
Frequency :	122	60	15	2	1

Graduate the data by Poisson distribution and calculate the theoretical frequencies.

Given	e^{-m} :	0·6703	0·6065	0·5488	0·4966
	m :	0·4	0·5	0·6	0·7

(c) Fit a Poisson distribution to the following data and calculate the expected frequencies :-

x :	0	1	2	3	4	5	6	7	8
f :	71	112	117	57	27	11	3	1	1

16. (*a*) If X is the number of occurrences of the Poisson variate with mean λ ; show that : $P(X \geq n) - P(X \geq n+1) = P(X = n)$

(*b*) Suppose that X has a Poisson distribution. If

$$P(X = 2) = \tfrac{2}{3}\, P(X = 1).$$

Evaluate (*i*) $P(X = 0)$ and (*ii*) $P(X = 3)$ [Ans. (*i*) 0·264.]

(c) If X has a Poisson distribution such that

$P(X = 1) = P(X = 2)$, find $P(X = 4)$. [Ans 0·09]

(c) If a Poisson variate X is such that

$$P(X = 1) = 2\,P(X = 2),$$

find $P\,(X = 0)$, mean and the variance.

(*d*) If for a Poisson variate X, $E(X^2) = 6$, what is $E(X)$?

(*e*) If X and Y are independent Poisson variates having means 1 and 3 respectively, find the variance of $3X + Y$.

17. Show that for a Poisson distribution

(*i*) $M\sigma\,\gamma_1\,\gamma_2 = 1$, (*ii*) $\beta_1^{1/2}\,(\beta_2 - 3)\,\mu_1'\sigma = 1$

18. Show that the function which generates the central moments of the Poisson distribution with parameter λ is

$$M(t) = \exp\left\{\lambda(e^t - 1 - t)\right\}$$

Show that it satisfies the equation

$$\frac{dM(t)}{dt} = \lambda t\,M(t) + \lambda\,\frac{dM(t)}{d\lambda}$$

19. (*a*) The random variable X has p.d.f.

$$f(x) = e^{-\theta}\,\frac{\theta^x}{x\,!};\ x = 0, 1, 2, \ldots$$
$$= 0, \text{ elsewhere}$$

Find the m.g.f. of $Y = 2X - 1$ and $\mathrm{Var}\,(Y)$.

(*b*) Identify the distribution with the following mgf's :

$$M_X(t) = (0{\cdot}3 + 0{\cdot}7\,e^t)^{10}$$
$$M_Y(t) = \exp\,[3(e^t - 1)]$$

Ans. $X \sim B\,(10, 0{\cdot}7)$, $Y \sim P(3)$.

20. If X has Poisson distribution with parameter λ, then

$$P\,[X \text{ is even}] = \frac{1}{2}\left[1 + e^{-2\lambda}\right]$$

[Delhi Univ. B. Sc. (Stat. Hons.) 1991]

21. (*a*) The m.g.f. of a r.v. is X is $\exp\,[4(e^t - 1)]$. Show that

$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{\cdot}931$

Hint. $X \sim P(\lambda = 4)$;

Required Probability. $= P(0 < X < 8) = P(1 \le X \le 7) = 0{\cdot}931$

(*b*) If $X \sim P(\lambda = 100)$, use Chebychev's inequality to determine a lower bound for $P(75 < X < 125)$ **[Ans. 0·84]**

22. If $X \sim P(m)$, show that $E\,|X - 1| = m - 1 + 2e^{-m}$

[Delhi Univ. B. Sc. (Maths. Hons.), 1983]

Hint. $E\,|X - 1| = \sum\limits_{x=0}^{\infty} |x - 1|\,e^{-m}\,m^x/x\,! = e^{-m} + \sum\limits_{x=2}^{\infty} \frac{(x-1)}{x\,!}\cdot e^{-m}\,m^x$

$$= e^{-m} + e^{-m}\cdot\sum_{x=2}^{\infty} m^x\left[\frac{1}{(x-1)\,!} - \frac{1}{x\,!}\right]$$

23. If $X \sim P(\lambda)$ and $Y|X = x \sim (B(x, p))$, then prove that $Y \sim P(\lambda p)$.

24. If the chances of 0, 1, 2, 3... events from one source are given by a Poisson distribution of mean m_1 and the chances of 0, 1, 2, 3,... events from another source by a Poisson distribution of mean m_2, show that the chances of 0, 1, 2, 3,... events from either source are given by

$$e^{-(m_1+m_2)}\left\{1,\ (m_1 + m_2),\ \frac{(m_1 + m_2)^2}{2!}, \ldots\right\}.$$

Show that the sum of any finite number of Poisson variates is itself a Poisson variate with mean equal to the sum of separate means.

25. X is a Poisson variate with mean λ.

Show that $E(X^2) = \lambda E(X + 1)$.

If $\lambda = 1$, show that $E|X - 1| = \dfrac{2}{e}$

26. Show that the mean deviation about mean for Poisson distribution

$$p(x) = \frac{e^{-m}\, m^x}{x!}\ ;\ x = 0, 1, 2, \ldots$$

is $(2\mu) \cdot \dfrac{e^{-m} \cdot m^{\mu}}{\mu!}$

where μ is the greatest integer contained in $(m + 1)$.

[Delhi Univ. B. Sc. (Stat. Hons.), 1988,' 84]

27. Let X, Y be independent Poisson variates. The variance of $X + Y$ is 9 and

$$P(X = 3 \,|\, X + Y = 6) = 5/54$$

Find the mean of X. [Ans. $\frac{1}{2}(9 \pm 3\sqrt{3})$ *i.e.* 1·902 *or* 7·098]

28. If X is a Poisson variate with paramter m, show that

$$P(X < r) < \frac{m^r}{r!}\ ;\ r = 0, 1, 2, \ldots$$

Deduce that $E(X) < e^m$. [Delhi Univ. B.Sc. (Maths. Hons.), 1989]

29. (*a*) The characteristic function of a variate X is

$$\varphi_X(t) = \left(\frac{1}{3} + \frac{2}{3}e^{it}\right)^6 \cdot \left[\exp\left\{-3(1 - e^{it})\right\}\right].$$

Recognise the variate.

[Burdwan Univ. B. Sc. (Maths. Hons.) 1989)

Hint. $X = U + V$, where $U \sim B\left(6, \frac{2}{3}\right)$ and $V \sim P(3)$ are independent *r.v.'s*

(*b*) Identify the variates X and Y where :

$M_X(t) = (1/27)(1 + 2e^t)^3 \cdot \exp\left[3(e^t - 1)\right]$

$$M_Y(t) = (1/32)(1 + e^t)^5 \cdot \exp\left[-2(1 - e^t)\right]$$

[Delhi Univ. B. Sc. (Stat. Hons.), 1987, 84]

Ans. $X = U + V$; $U \sim B(n = 3, p = 2/3)$ and $V \sim P(\lambda = 3)$ are independent.

$Y = U_1 + V_1$; $U_1 \sim B(n = 5, p = 1/2)$ and $V_1 \sim P(\lambda = 2)$ are independent.

30. If X and Y are correlated variates each having Poisson distribution, show that $X + Y$ cannot be a Poisson variate

[Delhi Univ. B. Sc. (Maths Hons.), 1988; Poona Univ. B.Sc., 1989]

Hint. Note that for Poisson variate mean and variance are equal. Let $X \sim P(\lambda)$, $Y \sim P(\mu)$; (X, Y) correlated.

$$\therefore\ E(X + Y) = E(X) + E(Y) = \lambda + \mu$$

$$\begin{aligned} \text{Var}(X + Y) &= \text{Var}\,X + \text{Var}\,Y + 2\,\text{Cov}(X, Y) \\ &= \lambda + \mu + 2\rho\sqrt{\lambda\mu},\ (\rho \neq 0) \end{aligned}$$

Since $E(X + Y) \neq \text{Var}(X + Y)$; $X + Y$ cannot be a Poisson variate.

31. Let X, Y, Z be independent Poisson variates with parameters a, b and c respectively. Obtain :

(*i*) m.g.f. of $X + 2Y + 3Z$,

(*ii*) Conditional expectation of X given $X + Y + Z = n$

(Indian Civil Services, 1985)

Hint.
$$\begin{aligned} M_{X+2Y+3Z}(t) &= M_X(t) \cdot M_Y(2t) \cdot M_Z(3t) \\ &= \exp\left[a\left(e^t - 1\right) + b\left(e^{2t} - 1\right) + c\left(e^{3t} - 1\right)\right] \end{aligned}$$

$$\begin{aligned} P(X = x \mid X + Y + Z = n) &= \frac{P(X = x \cap X + Y + Z = n)}{P(X + Y + Z = n)} \\ &= \frac{P(X = x)P(Y + Z = n - x)}{P(X + Y + Z = n)} \quad (\because X, Y, Z \text{ are indep.}) \\ &= \frac{e^{-a} \cdot a^x}{x!} \times \frac{e^{-(b+c)} \cdot (b + c)^{n-x}}{(n - x)!} \\ &\quad \times \left[\frac{n!}{e^{-(a+b+c)} \cdot (a + b + c)^n}\right] \\ &= \frac{n!}{x!(n - x)!}\left(\frac{a}{a + b + c}\right)^x \cdot \left(\frac{b + c}{a + b + c}\right)^{n} \end{aligned}$$

$$\Rightarrow \quad X \mid (X + Y + Z = n) \sim B\{n, p = a/(a + b + c)\}$$

$$\Rightarrow \quad E[X \mid X + Y + Z = n] = np = \frac{na}{a + b + c}$$

32. The joint density of r.v.'s X and Y is :

$$f(x, y) = e^{-2} / [x! \cdot (y - x)!] ;\ y = 0, 1, 2, \ldots ;\ x = 0, 1, 2, \ldots, y.$$

Find the m.g.f. $M(t_1, t_2)$ of (X, Y) and correlation coefficient between X and Y. Show that the marginal distributions of X and Y are Poisson.

Hint. $M(t_1, t_2) = \sum\limits_{y=0}^{\infty} \sum\limits_{x=0}^{y} e^{t_1 x + t_2 y} \times \left[\dfrac{e^{-2}}{x!(y-x)!} \right]$

$$= e^{-2} \sum_{y=0}^{\infty} \left[\frac{e^{t_2 y}}{y!} \left\{ \sum_{x=0}^{y} {}^yC_x \cdot (e^{t_1})^x \right\} \right]$$

$$= e^{-2} \sum_{y=0}^{\infty} \left\{ \left[e^{t_2} \left(1 + e^{t_1}\right) \right]^y / y! \right\}$$

$$= e^{-2} \cdot \exp \left\{ e^{t_2} (1 + e^{t_1}) \right\}$$

$$M(t_1, 0) = \exp\left[2(e^{t_1} - 1)\right] \Rightarrow X \sim P(\lambda = 1)$$

$$M(0, t_2) = \exp\left[2(e^{t_2} - 1)\right] \Rightarrow Y \sim P(\mu = 2)$$

Observe $M(t_1, t_2) \neq M(t_1, 0) \times M(0, t_2) \Rightarrow X$ and Y are not independent.

$E(X) = 1$, $\text{Var}(X) = 1$; $E(Y) = 2 = \text{Var}\, Y$.

$$E(XY) = \left| \frac{\partial^2 M(t_1, t_2)}{\partial t_1 \, \partial t_2} \right|_{t_1 = t_2 = 0} = 3$$

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{3 - 1 \times 2}{1 \times \sqrt{2}} = 1/\sqrt{2}.$$

33. The joint p.g.f. of the r.v.'s X and Y is given by :

$$P(s_1, s_2) = \exp\left[a(s_1 - 1) + b(s_2 - 1) + c(s_1 - 1)(s_2 - 1) \right],$$

a,b,c, are all positive. Find $\rho(X, Y)$

Hint. $P_X(s_1) = P(s_1, 1) = \exp[a(s_1 - 1)] \Rightarrow X \sim P(a)$

$P_Y(s_2) = P(1, s_2) = \exp[b(s_2 - 1)] \Rightarrow Y \sim P(b)$

$$E(XY) = \left(\frac{\partial^2 P(s_1, s_2)}{\partial s_1 \, \partial s_2} \right)_{s_1 = s_2 = 1} = c + ab.$$

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{(c + ab) - ab}{\sqrt{a}\sqrt{b}} = \frac{c}{\sqrt{ab}}$$

34. An insurance company issues only two types of policy, household and motor. It has carried out an investigation into the experience of a group of policyholders who held one of each type of policy over a particular period and it has discovered that within that group and over that period the mean number of claims per household policy was 0·3 and the mean number of claims per motor policy was 0·8. Assume that the number of claims under each type of policy is independent of the number of claims under the other type of policy and that each can be represented by a Poisson distribution.

(*a*) If the number of claims per policyholder is the sum of the number of claims under each of his two policies, state with reasons how the number of claims per policyholder, within that group and over that period is distributed, and (*b*) Calculate to the nearest whole number, the percentage of policyholders within that group and over that period who made more household claims than motor claims.

Hint. Household claim, $X \sim P(\cdot 3)$ and Motor claim, $Y \sim P(\cdot 8)$

$$\text{Required Probability} = P(X > Y) = \sum_{r=0}^{\infty}\left[\sum_{s=0}^{\infty} P(Y = r \cap X = r + s)\right]$$

$$= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\left[P(Y=r)P(X=r+s)\right] = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{e^{-\cdot 8}(\cdot 8)^r}{r!} \times \frac{e^{-\cdot 3}(\cdot 3)^{r+s}}{(r+s)!}$$

$$= e^{-\cdot 8}\, e^{-\cdot 3}\sum_{r=0}^{\infty}\left[\frac{(\cdot 8)^r}{r!}\sum_{s=0}^{\infty}\left\{\frac{(\cdot 3)^{r+s}}{(r+s)!}\right\}\right]$$

$$= \sum_{r=0}^{\infty}\left[\frac{e^{-\cdot 8}(\cdot 8)^r}{r!}\, e^{-\cdot 3}\left\{e^{\cdot 3} - \left(1 + \cdot 3 + \frac{(\cdot 3)^2}{2!} + \dots + \frac{(\cdot 3)^{r-1}}{(r-1)!}\right)\right\}\right]$$

$$= 1 - e^{-\cdot 8}e^{-\cdot 3}\left[\left\{\frac{\cdot 8}{1} + \frac{(\cdot 8)^2}{2!}(1 + \cdot 3) + \frac{(\cdot 8)^3}{3!}\left(1 + \cdot 3 + \frac{\cdot 09}{2}\right)\right.\right.$$

$$\left. + \frac{(\cdot 8)^4}{4!}\left(1 + \cdot 3 + \frac{\cdot 09}{2} + \frac{\cdot 027}{3!}\right) + \dots\right]$$

35. (*i*) An event occurs instantaneously and is equally likely to occur at any instant. There is no limit on the number of occurrneces that may happen in any interval of time, but the expected number in a given time interval is T. Prove that the probability of the event occurring exactly r times in an interval of the same duration is $(T^r\, e^{-T})/r\,!$.

(*ii*) An insurance company which writes only fire and accident business defines a major claim as one which costs at least Rs. 50,000 for an accident claim or Rs. 100,000 for a fire claim. Any excess over these amounts is paid by reinsurers and hence every major claim is recorded at a cost of Rs. 50,000 or Rs. 100,000 respectively. The company divides the year into equal monthly accounting periods and a report is produced of the recorded cost of major claims. The expected number of major accident claims is 0·2 per month and of major fire claims 0·5 per month. Calculate the probability that in a particular month the recorded cost of major claims is Rs. 2,00,000 or more.

36. (*a*) The number of aeroplanes arriving at an airport in a 30 minute interval obeys the Poisson law with mean 25. Use Chebychev's inequality to find the least chance, that the number of planes to arrive within a given 30 minutes interval will be between 15 and 35. **[Sri Venketeswara U. B.Sc. 1992]**

(*b*) Suppose that the number of motor cars arriving in a certain parking lot in any 15 minutes period obeys a Poisson probability law with mean 80. Use Chebychev's inequality to determine a lower bound for the probability that the number of motor cars arriving in a given 15 minute period will be between 60 and 100. [Madras U. B.Sc. Nov. 1991]

7·4. Negative Binomial Distribution. The equality of the mean and varaince is an important characteristic of the Poisson distribution, whereas for the binomial distribution the mean is always greater than the variance. Occasionally, however, observable phenomena give rise to empirical discrete distributions which show a variance larger than the mean. Some of the commonest examples of such behaviour are the frequency distributions of plant density obtained by quadrant sampling when the clustering of plants makes the simple Poisson model inapplicable. It has been shown by different investigators that in such cases the negative binomial distribution provides an excellent model because this distribution has a variance larger than the mean. Bacterial clustering (or contagion), *e.g.*, deaths of insects, number of insect bites leads to the negative binomial distribution and the distribution also arises in inverse sampling from a binomial population or as a weighted average of Poisson distribution. This important probability distribution is sometimes also referred to as the Pascal distribution after the French mathematician Blaise Pascal (1623-1662), but there seems to be no historical justification. The negative binomial distribution can be derived from empirical considerations in many ways. Here we consider the Binomial probability situation with some modifications.

Suppose we have a succession of n Bernoulli trails. We assume that (*i*) the trials are independent, (*ii*) the probability of success 'p' in a trial remains constant from trial to trial.

Let $f(x\,;r,p)$ denote the probability that there are x failures preceding the rth success in $x+r$ trials.

Now, the last trial must be a success, whose probability is p. In the remaining $(x+r-1)$ trials we must have $(r-1)$ successes whose probability is given by

$$\binom{x+r-1}{r-1} p^{r-1} q^x \quad .$$

Therefore by compound probability theorem, $f(x\,;r,p)$ is given by the product of these two probabilities, *i.e.*,

$$\binom{x+r-1}{r-1} p^{r-1} q^x \cdot p = \binom{x+r-1}{r-1} p^r q^x$$

Definition. *A random variable X is said to follow a negative binomial distribution if its probability mass function is given by*

$$p(x) = P(X=x) = \binom{x+r-1}{r-1} p^r q^x\,;\; x = 0, 1, 2, \ldots$$
$$= 0, \text{ otherwise} \qquad \ldots(7\cdot21)$$

Also

$$\binom{x+r-1}{r-1} = \binom{x+r-1}{x} \qquad \left[\because \binom{n}{r} = \binom{n}{n-r}\right]$$

$$= \frac{(x+r-1)(x+r-2)\ldots(r+1)r}{x!}$$

$$= \frac{(-1)^x(-r)(-r-1)\ldots(-r-x+2)(-r-x+1)}{x!}$$

$$= (-1)^x \binom{-r}{x}$$

$$\therefore \qquad p(x) = \begin{cases} \binom{-r}{x} p^r(-q)^x; x = 0, 1, 2, \ldots \\ 0, \text{ otherwise} \end{cases} \qquad \ldots(7\cdot21\ a)$$

which is the $(x+1)^{th}$ term in the expansion of $p^r(1-q)^{-r}$, a binomial expansion with a negative index. Hence the distribution is known as negative binomial distribution. Also

$$\sum_{x=0}^{\infty} p(x) = p^r \sum_{x=0}^{\infty} \binom{-r}{x} (-q)^x = p^r \times (1-q)^{-r} = 1$$

Therefore $p(x)$ represents the probability function and the discrete variable which follows this probability function is called the negative binomial variate.

If $\qquad p = \dfrac{1}{Q}$ and $q = \dfrac{P}{Q}$ so that $Q - P = 1$, $\quad (\because p + q = 1)$

$$\text{then} \quad p(x) = \begin{cases} \binom{-r}{x} Q^{-r} \left(-\dfrac{P}{Q}\right)^x ; x = 0, 1, 2, \ldots \\ 0, \text{ otherwise} \end{cases} \qquad \ldots(7\cdot21\ b)$$

This is the general term in the negative binomial expansion $(Q-P)^{-r}$.

Remarks. 1. $p(x)$ in (7·21) or (7·21*a*) is also sometimes written as $f(x; r, p)$.

2. Some Important Deductions.

(*a*) **Geometric Distribution.** If we take $r = 1$ in (7·21), we have

$$p(x) = q^x p; \ x = 0, 1, 2, \ldots$$

which is the probability function of geometric distribution (*c.f.* § 7·5 page 7·83).

Hence negative binomial distribution may be regarded as the generalisation of geometric distribution.

(*b*) **Pascal's Distribution.** The negative binomial distribution (7·21 *a*) when regarded as one having two parameters p and r is known as *Pascal's distribution.*

(*c*) **Polya's Distribution.** If we take

$r = \dfrac{1}{\beta}$, $p = \dfrac{1}{1 + \beta\mu}$, $q = 1 - p = \dfrac{\beta\mu}{1 + \beta\mu}$ in (7·21 *a*), we get

$$p(x) = \frac{r(r+1)(r+2)\ldots(r+x-1)}{x!} \cdot p^r \cdot q^x$$

$$= \frac{(1+\beta)(1+2\beta)\dots[1+\beta(x-1)]}{x!}\left(\frac{1}{1+\beta\mu}\right)^{1/\beta}\left(\frac{\mu}{1+\beta\mu}\right)^{x}$$

$$(x = 0, 1, 2, \dots) \qquad \dots(7{\cdot}21\ c)$$

which is known as *Polya's distribution* with two parameters, β and μ.

(*d*) Second Form of Geometric Distribution. Taking $\beta = 1$ in Polya's distribution (7·21 *c*), we get

$$p(x) = \left(\frac{1}{1+\mu}\right)\left(\frac{\mu}{1+\mu}\right)^{x}; \; x = 0, 1, 2, \dots \qquad \dots(7{\cdot}21\ d)$$

which is geometric distribution (c.f. § 7·5) with

$$p = \frac{1}{1+\mu}, \; q = 1 - p = \frac{\mu}{1+\mu}$$

7·4·1. Moment Generating Function of Negative Binomial Distribution.

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx}\, p(x)$$

$$= \sum_{x=0}^{\infty}\binom{-r}{x} Q^{-r}\left(-\frac{Pe^t}{Q}\right)^{x}$$

$$= (Q - Pe^t)^{-r} \qquad \dots(7{\cdot}22)$$

$$\mu_1' = \left(\frac{d}{dt}M(t)\right)_{t=0} = \left[-r(-Pe^t)(Q - Pe^t)^{-r-1}\right]_{t=0}$$

$$= rP$$

$\therefore$ Mean of the negative binomial distribution is rP.

$$\mu_2' = \left(\frac{d^2}{dt^2}M(t)\right)_{t=0} \qquad \dots(7{\cdot}22\ a)$$

$$= \left(rPe^t(Q - Pe^t)^{-r-1} + (-r-1)rPe^t(Q - Pe^t)^{-r-2}(-Pe^t)\right)_{t=0}$$

$$= rP + r(r+1)P^2$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = r(r+1)P^2 + rP - r^2p^2 = rPQ \qquad \dots(7{\cdot}22\ b)$$

As $Q > 1$, $rP < rPQ$, *i.e., Mean < Variance, which is a distinguishing feature of this distribution.*

7·4·2. Cumulants of Negative Binomial Distribution.

$$K_X(t) = \log M_X(t) = -r\log(Q - Pe^t)$$

$$= -r\log\left[Q - P\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \dots\right)\right]$$

$$= -r\log\left[1 - P\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \dots\right)\right]$$

$$(\because Q - P = 1)$$

Proceeding as in § 7·2·8, we will get (on replacing n with $-r$ and p with $-P$).

$$\text{Mean} = \kappa_1 = rP$$
$$\mu_2 = \kappa_2 = r\,P(1+P) = rPQ \qquad \ldots(7\cdot23)$$
$$\mu_3 = \kappa_3 = rP(1+3P+2P^2) = rP(1+P)(1+2P) = rPQ(Q+P).$$
$$\kappa_4 = rP(1+P)(1+6P+6P^2) = rPQ(1+6PQ)$$
$$\therefore \quad \mu_4 = \kappa_4 + 3\kappa_2^2 = rPQ[1+3PQ(r+2)]$$

Since $Q = 1/p$, $P = qQ = q/p$, we have in terms of p and q,

Mean $= rq/p$, Variance $= rq/p^2$, $\mu_3 = rq(1+q)/p^3$

$$\mu_4 = rq\,[p^2 + 3q(r+2)]/p^4$$
$$\therefore \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(1+q)^2}{rq}$$
$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{p^2+3q(r+2)}{rq} \qquad \ldots(7\cdot23\,a)$$
$$\therefore \quad \gamma_1 = \sqrt{\beta_1} = (1+q)/\sqrt{rq}$$
$$\gamma_2 = \beta_2 - 3 = (p^2+6q)/rq$$

7·4·3. Poisson Distribution as a Limiting Case of the Negative Binomial Distribution. Negative binomial distribution tends to Poisson distribution as $P \to 0$, $r \to \infty$ such that $rP = \lambda$ (finite). Proceeding to the limits, we get

$$\lim p(x) = \lim \binom{x+r-1}{r-1} p^r q^x$$
$$= \lim \binom{x+r-1}{x} Q^{-r} \left(\frac{P}{Q}\right)^x$$
$$= \lim_{r\to\infty} \frac{(x+r-1)(x+r-2)\ldots(r+1)r}{x!}(1+P)^{-r}\left(\frac{P}{1+P}\right)^x$$
$$= \lim_{r\to\infty}\left[\frac{1}{x!}\left(1+\frac{x-1}{r}\right)\left(1+\frac{x-2}{r}\right)\ldots\left(1+\frac{1}{r}\right).1.r^x(1+P)^{-r}\left(\frac{P}{1+P}\right)^x\right]$$
$$= \frac{1}{x!}\lim_{r\to\infty}\left[(1+P)^{-r}\left(\frac{rP}{1+P}\right)^x\right]$$
$$= \frac{\lambda^x}{x!}\lim_{r\to\infty}\left[\left(1+\frac{\lambda}{r}\right)\right]^{-r}\lim_{r\to\infty}\left(1+\frac{\lambda}{r}\right)^{-x} \qquad [\because rP = \lambda]$$
$$= \frac{\lambda^x}{x!}\cdot e^{-\lambda}\cdot 1 = \frac{e^{-\lambda}\lambda^x}{x!}$$

which is the probability function of the Poisson distribution with parameter 'λ'.

7·4·4. Probability Generating Function of Negative Binomial Distribution. Let X be a random variable following negative binomial distribution, then

$$P_X(s) = E(s^X) = \sum_{x=0}^{\infty} s^x \, p(x)$$

$$= \sum_{x=0}^{\infty} \binom{-r}{x} p^r (-qs)^x \qquad \text{[Using 7·21 a)]}$$

$$= p^r (1 - qs)^{-r} = [p/(1 - qs)]^r \qquad \ldots(7·24)$$

Example 7·45. *An item is produced in large numbers. The machine is known to produce 5% defectives. A quality control inspector is examining the items by taking them at random. What is the probability that at least 4 items are to be examined in order to get 2 defectives ?*

Solution. If 2 defectives are to be obtained then it can happen in 2 or more trials. The probability of success is 0·05 for every trial. It is a negative binomial situation and the required probability is

$$= P(X = 4) + P(X = 5) + \ldots$$

$$= \sum_{x=4}^{\infty} \binom{x-1}{2-1} (0·05)^2 (0·95)^{x-2}$$

$$= 1 - \sum_{2}^{3} \binom{x-1}{2-1} (0·05)^2 (0·95)^{x-2}$$

$$= 1 - [(0·05)^2 + 2(0·05)^2(0·95)]$$

$$= 0·995$$

Example 7·46. *If $X \sim B(n, p)$ and Y has negative binomial distribution with parameters r and p, prove that*

$$F_X(r-1) = 1 - F_Y(n-r)$$

[Delhi Univ. Spl. Course (Statistics Hons.), 1987]

Solution.

$$1 - F_Y(n-r) = 1 - P(Y \le n-r) = P(Y > n-r)$$

$$= \sum_{n-r+1}^{\infty} \binom{y+r-1}{r-1} p^r \cdot q^y ; \; [z = y - (n-r+1)]$$

$$= p^r q^{n-r+1} \cdot \sum_{z=0}^{\infty} \binom{z+n}{r-1} q^z$$

$$= p^r q^{n-r+1} \sum_{z=0}^{\infty} \left\{ \sum_{k=0}^{r-1} \binom{n}{k} \binom{z}{r-1-k} \right\} q^z$$

$$\left(\because \sum_{k=0}^{r}\binom{a}{k}\binom{b}{r-k}=\binom{a+b}{r}\right)$$

$$= p^r q^{n-r+1}\sum_{k=0}^{r-1}\left\{\binom{n}{k}\sum_{z=r-1-k}^{\infty}\binom{z}{r-1-k}q^z\right\}$$

[Changing the order of summation and noting that $\binom{n}{r}=0;\ n<r$]

$$= p^r q^{n-r+1}\sum_{k=0}^{r-1}\left[\binom{n}{k}\sum_{t=0}^{\infty}\binom{t+r-1-k}{r-1-k}q^{t+r-1-k}\right],$$

$$t = z-(r-1-k)$$

$$= p^r q^n\sum_{k=0}^{r-1}\left\{\binom{n}{k}q^{-k}\cdot(1-q)^{-(r-k)}\right\}$$

$$= \sum_{k=0}^{r-1}\binom{n}{k}p^k\cdot q^{n-k}$$

$$= P[X \le (r-1)] = F_X(r-1)$$

Example 7·47. (Banach's Match-box Problem). *A certain mathematician always carries two match boxes (initially containing N match sticks). Each time he wants a match-stick he selects a box at random, inevitably a moment comes when he finds a box empty. Show that the probability that there are exactly r match-sticks in one box when the other box is found empty is*

$$\binom{2N-r}{n}\times\left(\frac{1}{2}\right)^{2N-r}$$

Solution. Let the two match boxes be numbered 1 and 2. Let the choice of the Ist box be regarded as failure and that of second box be regarded as a success.

Since the mathematician selects the match box at random,

p = Probability of selecting second match box = ½

$\Rightarrow \qquad q = 1 - p = ½$

The second box will be found empty if it is selected for the $(N+1)$st time. At this stage, the first box will contain exactly r matches if $(N-r)$ matches have already been drawn from it. Hence the second box will be found empty at the stage when the first box contains exactly r matches if and only if $(N-r)$ failures precede the $(N+1)$st success. Thus in a total of $N+1+(N-r) = 2N-r+1$, trials the last one must be success and out of the remaining $(2N-r)$ trials we should have $(N-r)$ failures and N successes.

$\therefore$ Probability that second box is found empty when there are exactly r matches in first box is

$$= \binom{2N-r}{N}\left(\frac{1}{2}\right)^{N}\left(\frac{1}{2}\right)^{N-r}\frac{1}{2} = \binom{2N-r}{N}\left(\frac{1}{2}\right)^{2N-r+1}$$

Similarly, the probability that first box is found empty, when the second box contains exactly r matches is given by

$$\binom{2N-r}{N}\left(\frac{1}{2}\right)^{2N-r+1}$$

Hence the required probability that one match box is found empty when the other contains exactly r matches is

$$2 \times \binom{2N-r}{N}\left(\frac{1}{2}\right)^{2N-r+1} = \binom{2N-r}{N}\left(\frac{1}{2}\right)^{2N-r}$$

Remark. The statement that 'he finds the box empty' implies that when he used the last match in this box, he did not throw it away, but instead put it back in his pocket. Thus there is a difference between 'the box is empty' and 'the box is found empty'.

The box becomes empty when the Nth match was taken from it but it is found to be empty only when it is selected for the $(N+1)$st time.

Example 7·48. *X is a negative binomial variate with p.f.*

$$f(x) = \begin{cases} \binom{k+x-1}{x} q^x p^k, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

Show that the moment recurrence formula is

$$\mu_{r+1} = q\left[\frac{d\mu_r}{dq} + \frac{rk}{p^2}\mu_{r-1}\right]$$

State how the moments of negative binomial variate can be written from the corresponding formulas for binomial variate.

[Punjab Univ. B. Sc. (Maths Hons.) 1990]

Solution. For Negative Binomial Distribution with parameter k and p,

$$\text{Mean} = k \cdot q/p = \mu, \text{ (say)}.$$

$$\therefore \quad \mu_r = \sum_{x=0}^{\infty} (x-\mu)^r f(x)$$

$$= \sum_{x=0}^{\infty}\left[\left(x-\frac{kq}{p}\right)^r \cdot \binom{k+x-1}{x} q^x \cdot p^k\right]$$

Differentiating w.r.to q, we get

$$\frac{d\mu_r}{dq} = \sum_x\left[r\left(x-\frac{kq}{p}\right)^{r-1} \times \left\{\frac{d}{dq}\left(x-\frac{kq}{p}\right)\right\}\binom{k+x-1}{x} q^x p^k\right]$$

$$+ \sum_x\left[\left(x-\frac{kq}{p}\right)^r \binom{k+x-1}{x}\left\{xq^{x-1}p^k + q^x \cdot kp^{k-1} \cdot \frac{dp}{dq}\right\}\right]$$

But $\dfrac{dp}{dq} = \dfrac{d}{dq}(1 - q) = -1$

and $\dfrac{d}{dq}\left[x - \dfrac{kq}{p}\right] = \dfrac{d}{dq}\left[x - k\left(\dfrac{1}{p} - 1\right)\right] = \dfrac{k}{p^2} \cdot \dfrac{dp}{dq} = -\dfrac{k}{p^2}$

$$\therefore \quad \frac{d\mu_r}{dq} = -\frac{rk}{p^2}\sum_x \left(x - \frac{kq}{p}\right)^{r-1} \cdot f(x)$$

$$+ \sum_x \left(x - \frac{kq}{p}\right)^r \binom{k + x - 1}{x} q^{x-1} p^k \left(x - \frac{kq}{p}\right)$$

$$= -\frac{rk}{p^2}\mu_{r-1} + \frac{1}{q}\sum_x \left(x - \frac{kq}{p}\right)^{r+1} \cdot f(x)$$

$$= -\frac{rk}{p^2}\mu_{r-1} + \frac{1}{q} \cdot \mu_{r+1}$$

$$\Rightarrow \quad \mu_{r+1} = q\left[\frac{d\mu_r}{dq} + \frac{rk}{p^2}\mu_{r-1}\right];\ r = 1, 2, 3, \ldots$$

7·4·5. Deduction of Moments of Negative Binomial Distribution From Those of Binomial Distribution. If we write

$$p = 1/Q,\ q = P/Q \text{ such that } Q - P = 1,$$

then the m.g.f. of negative binomial variate X is given by [c.f. § 7·4·1] :

$$M_X(t) = (Q - Pe^t)^{-k} \qquad \ldots(*)$$

This is analogous to the m.g.f. of binomial variate Y with parameters n and p', *viz.*,

$$M_Y(t) = (q' + p'e^t)^n;\ q' = 1 - p' \qquad \ldots(**)$$

Comparing (*) and (**), we get

$$q' = Q,\ p' = -P \text{ and } n = -k \qquad \ldots(***)$$

Using the formulae for moments of binomial distribution, the moments of negative binomial distribution are given by

Mean $= np' = (-k)(-P) = kP$

Variance $= np'q' = (-k)(-P)Q = kPQ$

$\mu_3 = np'q'(q' - p') = (-k)(-P)Q(Q + P) = kPQ(Q + P)$

$\mu_4 = np'q'[1 + 3p'q'(n - 2)]$

$= (-k)(-P)Q[1 + 3(-P)Q(-k - 2)]$

$= kPQ[1 + 3PQ(k + 2)]$

Example 7·49. *Prove that the recurrence formula for negative binomial distribution is :* $f(x + 1; r, p) = \dfrac{x + r}{x + 1}\, q f(x; r, p)$

(Utkal Univ. M.A., 1990)

Solution. We have

$$f(x\,;r,p) = \binom{x+r-1}{r-1} p^r q^x$$

$$f(x+1\,;r,p) = \binom{x+r}{r-1} p^r q^{x+1}$$

$$\therefore \quad \frac{f(x+1\,;r,p)}{f(x\,;r,p)} = \frac{(x+r)\,!\,(r-1)\,!\,x\,!}{(r-1)\,!\,(x+1)\,!\,(x+r-1)\,!}\,q = \frac{x+r}{x+1}\,q$$

$$\Rightarrow \quad f(x+1\,;r,p) = \frac{x+r}{x+1} \cdot q \cdot f(x\,;r,p)$$

This recurrence relation is useful for fitting of the negative binomial distribution to the given data as discussed in the following example.

Example 7·50. *Given the hypothetical distribution :*

No. of cells : (x)	*0*	*1*	*2*	*3*	*4*	*5*	*Total*
Frequency : (f)	*213*	*128*	*37*	*18*	*3*	*1*	*400*

Fit a negative binomial distribution and calculate the expected frequencies.

Solution. $\mu_1' = \text{Mean} = \dfrac{\Sigma fx}{\Sigma f} = \dfrac{473}{400} = \cdot 6825 = \dfrac{r.q}{p}$ ...(*)

$$\mu_2' = \frac{511}{400} = 1{\cdot}2775,$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 1{\cdot}2775 - (\cdot 6825)^2 = 0{\cdot}8117$$

$\therefore$ Variance $= 0{\cdot}8117 = \dfrac{rq}{p^2}$ ...(**)

Solving equations (*) and (**), we get

$$p = \frac{0{\cdot}6825}{0{\cdot}8117} = 0{\cdot}8408, \quad q = 1 - p = 0{\cdot}1592$$

$$\therefore \quad r = \frac{p \times 0{\cdot}6825}{q} = \frac{0{\cdot}5738}{0{\cdot}1592} = 3{\cdot}60456$$

$$f_0 = p^r = (\cdot 8408)^{3\cdot 6045} = 5352$$

$$f_1 = \frac{r+0}{0+1}\,q f_0 = rq f_0 = 0{\cdot}5738 \times 0{\cdot}5352 = 0{\cdot}3071$$

$$f_2 = \frac{r+1}{1+1} \cdot q \cdot f_1 = \frac{4{\cdot}60456}{2} \times 0{\cdot}1592 \times 0{\cdot}3071 = 0{\cdot}1126$$

$$f_3 = \frac{r+2}{2+1} \cdot q \cdot f_2 = \frac{5{\cdot}60456}{3} \times 0{\cdot}1592 \times 0{\cdot}1126 = 0{\cdot}0335$$

$$f_4 = \frac{r+3}{3+1} \cdot q \cdot f_3 = \frac{6{\cdot}60456}{4} \times 0{\cdot}1592 \times 0{\cdot}0335 = 0{\cdot}0088$$

$$f_5 = \frac{r+4}{4+1} \cdot q \cdot f_4 = \frac{7{\cdot}60456}{5} \times 0{\cdot}1592 \times 0{\cdot}0088 = 0{\cdot}000213$$

$\therefore$ Expected frequencies are :

Nf_0	Nf_1	Nf_2	Nf_3	Nf_4	Nf_5
214·0992	122·8596	45·0308	13·3928	3·5204	0·8524

$\therefore$ *Observed Frequency* :	213	128	37	18	3	1
Expected Frequency :	214	123	45	13	4	1

EXERCISE 7 (c)

1. (*a*) Define negative binomial distribution. Give an example in which it occurs. Obtain its moment generating function. Hence or otherwise obtain its mean, variance and third central moment. **[Gujarat Univ. B.Sc. 1992]**

(*b*) If X denotes the number of failures preceding the rth success in an infinite series of independent trials with constant probability p of success for each trial, then identify the distribution of X and obtain $E(X)$. What is the distribution when $r = 1$? **[Delhi Univ. B. Sc. (Stat. Hons.), 1985]**

2. (*a*). A well known baseball player has a lifetime batting average of 0·3. He needs 32 more hits to make up his lifetime total to 3000. What is the probability that 100 or fewer times at bat are required for him to achieve his goal?

(*b*) A scientist needs three diseased rabbits for an experiment. He has 20 rabbits availlable and inoculates them one at a time with a serum, quitting if and when he gets 3 positive reactions. If the probability is 0·25 that a rabbit can contract the disease from the serum, what is the probability that the scientist is able to get 3 diseased rabbits from 20 ?

3. A student has taken a 5 answer multiple choice examination orally. He continues to answer questions until he gets five correct answers. What is the probability that he gets them on or before the twenty-fifth question if he guesses at each answer ?

4. If a boy is throwing stones at a target, what is the probability that his 10th throw is his 5th hit, if the probability of hitting the target at any trial is 0·5 ?

5. In a series of independent trials with constant probability p of success in each trial, show that the number of successes in a fixed number n of independent trials follows a binomial distribution. Show further that the number of the trials required for a specified number r of successes follows a negative binomial distribution. Obtain the mean and the variance of this distribution.

6. (*a*) Obtain the Poisson-distribution as a limiting case of the negative binomial distribution. **[(Delhi Univ. B.Sc. (Stat Hons.) 1988]**

(*b*) Show how the moments of negative binomial variate can be written from the corresponding formulae for the bibomial variate.

[Delhi Univ. B.Sc. (Maths Hons.), 1991]

7. Consider a sequence of Bernoulli trials with constant probability p of success in a single trial. Find $P(x, k)$, the probability that exactly $x + k$ trials are required to get k successes, $x = 0, 1, 2, \ldots$ Show that $P(x, k)$ defines the probability function of the discrete random variable X. Find the moment generating function of X. Hence find $E(X)$ and $V(X)$.

8. (*a*) Derive moment generating function of negative binomial distribution and hence show that mean < variance

(*b*) Derive negative binomial distribution in the following form :

$$f(x) = \binom{-k}{x}(-P)^x Q^{-k-x}\,;\ \begin{matrix} x = 0, 1, 2, \ldots \\ Q = 1 + P \end{matrix}$$

Obtain (*i*) moment generating function, mean and variance of this distribution.

(*ii*) Coefficient of skewness β_1.

(*iii*) Give an example of its occurrence. [Gujarat Univ. B.Sc. Oct. 1990)

9. Obtain the characteristic funciion of the negative binomial distribution given in the form:

$$f(x\,;\alpha,\lambda) = \binom{-\lambda}{x}\left(\frac{\alpha}{1+\alpha}\right)^{\lambda}\left(\frac{-1}{1+\alpha}\right)^{x};\ x = 0, 1, 2,\ldots$$

and hence evaluate its first two moments.

10. (*a*) Show that for the negative binomial distribution $(Q - P)^{-r}$, where $Q - P = 1$, cumulant generating function $K(t) = -r\log[1 - P(e^t - 1)]$. Hence deduce that $\kappa_1 = rP$, $\kappa_2 = rPQ$ Also obtain κ_3 and κ_4.

[Delhi Univ. B.Sc. (Stat. Hons.) 1986]

(*b*) Show that the mean deviation about mean for the negative binomial distribution is

$$2(\mu + 1)\binom{n+\mu}{\mu+1}p^{\mu+1}\,q^{-(n+\mu)}$$

where μ is the greatest integer contained in $np + l$.

11. The number of accidents among 414 machine operators was investigated for three successive months. The following table gives the distribution of the operators according to the number k, of accidents which happened to the same operator. Fit the distribution of the type

$$P(X = k) = (-1)^k\binom{-v}{k}p^v q^k\,;\ k = 0, 1, 2,\ldots, v > 0, q = 1 - p, 0 < p < 1$$

k	...	0	1	2	3	4	5	6	7	8
Observed frequency	...	296	74	26	8	4	4	1	0	1

12. If X has negative binomial distribution with parameters (n, P), prove that $M_X(t) = (Q - Pe^t)^{-n}$. Hence find m.g.f. of $Z = (X - nP)/\sqrt{nPQ}$ and deduce that Z is asymptotically normal as $n \to \infty$

Hint. Prove that $M_Z(t) \to \exp(t^2/2)$ as $n \to \infty$ [c.f. Example 7·19].

13. Let Y have the negative binomial distribution. Let X_j be the number of failures between the $(j - 1)$th and jth success. Then $\sum_{j=1}^{r} X_j = Y$. Find $E(Y)$, σ_Y^2 by obtaining the means and variances of the X_j's.

14. Assume that the mutually independent random variables X_i, each have the negative binomial distribution with parameters r_i ($i = 1, 2, ..., n$), where r_i are all positive integers, i.e.,

$$P(X_i = x) = \binom{r_i + x - 1}{x} p^{r_i} q^x ; x = 0, 1, 2, ...$$

Then show that the probability density function of $\sum_{i=1}^{n} X_i$ is the negative binomial distribution with $r = \sum_{i=1}^{n} r_i$ *i.e.,* the negative binomial distribution (with fixed p) is reproductive with respect to r. **(Sagar Univ. M.A., 1991)**

15. Suppose that a radio tube is inserted into a socket and tested. Assume that the probability that it tests positive equals P and the probability that it tests negative is $(1 - P)$. Assume furthermore that we are testing large supply of such tubes. The testing continues until the first positive tube appears. If X is the number of tests required to terminate the experiment, what is the probability distribution of X ? **[(Aligarh U. B.Sc. (Hons.) 1993)]**

16. A man buys two boxes of matches, each containing N matches initially and places one match box in his right pocket and one in his left pocket. Every time when he wants a match, he selects a pocket at random. Show that the probability that at the moment when the first box is emptied (not found empty), the other box contain exactly r matches ($r = 1, 2, ..., N$) is

$$\binom{2N - 1 - r}{r - 1} \left(\frac{1}{2}\right)^{2N - r - 1}$$

Using this result, show that the probability that the box first emptied is not the one first found to be empty is

$$\left(\frac{1}{2}\right)^{2N} \cdot \sum_{r=1}^{N} \binom{2N - 1 - r}{N - 1},$$

which reduces to $\binom{2N}{N} \left(\frac{1}{2}\right)^{2N + 1}$ or $\frac{1}{2} (N\pi)^{-1/2}$ approximately.

7·5. Geometric Distribution. Suppose we have a series of independent trials or repetitions and on each repetition or trial the probability of success 'p' remains the same. Then the probability that there are x failures preceding the first success is given by $q^x p$.

Definition. *A random variable X is said to have a geometric distribution if it assumes only non-negative values and its probability mass function is given by*

$$P(X = x) = \begin{cases} q^x p ; x = 0, 1, 2, ..., 0 < p \le 1 \\ 0, \text{ otherwise} \end{cases} \quad ...(7·25)$$

Remarks. 1. Since the various probabilities for $x = 0, 1, 2, ...,$ are the various terms of geometric progression, hence the name geometric distribution.

2. Clearly, assignment of probabilities in (7·25) is permissible, since

$$\sum_{x=0}^{\infty} P(X=x) = \sum_{x=0}^{\infty} q^x p = p(1+q+q^2+\dots) = \frac{p}{1-q} = 1$$

7·5·1. Lack of Memory. The geometric distribution is said to lack memory in a certain sense. Suppose an event E can occur at one of the times $t = 0, 1, 2, \dots$ and the occurrence (waiting) time X has a geometric distribution. Thus $P(X=t) = q^t \cdot p;\ t = 0, 1, 2, \dots$

Suppose we know that the event E has not occurred before k, *i.e.*, $X \geq k$. Let $Y = X - k$. Thus Y is the amount of additional time needed for E to occur. We can show that

$$P(Y=t \mid X \geq k) = P(X=t) = pq^t \qquad \dots(7·26)$$

which implies that the additional time to wait has the same distribution as initial time to wait.

Since the distribution does not depend upon k, it, in a sense, '*lacks memory*' of how much we shifted the time origin. If 'B' were waiting for the event E and is relieved by 'C' immediately before time k, then the waiting time distribution of 'C' is the same as that of 'B'.

Proof of (7·26). We have

$$P(X \geq r) = \sum_{s=r}^{\infty} pq^s = p(q^r + q^{r+1} + \dots) = \frac{pq^r}{(1-q)} = q^r$$

$$P(Y \geq t \mid X \geq k) = \frac{P(Y \geq t \cap X \geq k)}{P(X \geq k)} = \frac{P(X-k \geq t \cap X \geq k)}{P(X \geq k)}$$

$$(\because Y = X - k)$$

$$= \frac{P(X \geq k+t)}{P(X \geq k)} = \frac{q^{t+k}}{q^k} = q^t$$

$$\therefore P(Y=t \mid X \geq k) = P(Y \geq t \mid X \geq k) - P(Y \geq t+1 \mid X \geq k)$$

$$= q^t - q^{t+1} = q^t(1-q) = pq^t = P(X=t)$$

7·5·2. Moments of Geometric Distribution.

$$\mu_1' = \sum_{x=1}^{\infty} x \cdot P(X=x) = \sum_{x=1}^{\infty} x \cdot pq^x = pq \sum_{x=1}^{\infty} xq^{x-1} = pq(1-q)^{-2} = \frac{q}{p}$$

$$V(X) = E(X^2) - [E(X)]^2 = E[X(X-1)] + E(X) - [E(X)]^2$$

$$E[(X-1)X] = \sum_{x=1}^{\infty} x(x-1)P(X=x) = \sum_{x=2}^{\infty} x(x-1)pq^x$$

$$= 2pq^2 \sum_{x=2}^{\infty} \left[\frac{x(x-1)}{2 \times 1} q^{x-2}\right] = 2pq^2(1-q)^{-3} = \frac{2q^2}{p^2}$$

$$\therefore \quad V(X) = \mu_2 = \frac{2q^2}{p^2} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q^2}{p^2} + \frac{q}{p} = \frac{q}{p^2}$$

7·5·3. Moment Generating Function of Geometric Distribution.

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \cdot q^x p = p \sum_{x=0}^{\infty} (e^t q)^x = p(1 - qe^t)^{-1}$$

$$= p/(1 - qe^t) \qquad \text{...(7·27)}$$

$$\mu_1' = \left[\frac{d}{dt} M(t)\right]_{t=0} = \left[\frac{d}{dt} p(1 - qe^t)^{-1}\right]_{t=0}$$

$$= p\,[qe^t(1 - qe^t)^{-2}]_{t=0} = pq(1 - q)^{-2} = \frac{q}{p}$$

$$\mu_2' = \left[\frac{d^2}{dt^2} M(t)\right]_{t=0} = \frac{q}{p} + \frac{2q^2}{p^2} \qquad \text{(On simplification)}$$

$$\mu_2 = \mu_2' - \mu_1'^2 = \frac{q}{p} + \frac{2q^2}{p^2} - \frac{q^2}{p^2} = \frac{q^2 + pq}{p^2} = \frac{q}{p^2}$$

Hence the mean and variance of the geometric distribution are q/p and q/p^2 respectively.

Remark. The p.g.f. of the geometric distribution is obtained on replacing e^t by s in (7·27) and is given by :

$$P_X(s) = p/(1 - qs) \qquad \text{...(7·27 a)}$$

Example 7·51. *Let the two independent random variables X_1 and X_2 have the same geometric distribution. Show that the conditional distribution of $X_1 | (X_1 + X_2 = n)$ is uniform.*

[Gujarat Univ. B.Sc. 1992; Calicut U. B.Sc. (Main Stat), Oct. 1990]

Solution. We are given

$$P(X_1 = k) = P(X_2 = k) = pq^k \,;\, k = 0, 1, 2 \ldots$$

$$P[X_1 = r | (X_1 + X_2 = n)] = \frac{P(X_1 = r \cap X_1 + X_2 = n)}{P(X_1 + X_2 = n)}$$

$$= \frac{P(X_1 = r \cap X_2 = n - r)}{P(X_1 + X_2 = n)}$$

$$= \frac{P(X_1 = r \cap X_2 = n - r)}{\sum_{s=0}^{n} [P(X_1 = s) \cap X_2 = n - s)}$$

$$= \frac{P(X_1 = r) \cdot P(X_2 = n - r)}{\sum_{s=0}^{n} [P(X_1 = s) \cdot P(X_2 = n - s)]}$$

[Since X_1 and X_2 are independent)

$$\therefore \quad P[X_1 = r | (X_1 + X_2 = n)] = \frac{pq^r \cdot pq^{n-r}}{\sum_{s=0}^{n} [pq^s \cdot pq^{n-s}]} = \frac{p^2 q^n}{\sum_{s=0}^{n} (p^2 q^n)}$$

$$= \frac{p^2 q^n}{(n+1)p^2 q^n} = \frac{1}{n+1};\ r = 0, 1, 2, \ldots n$$

Hence the conditional distribution of $X_1 \mid (X_1 + X_2 = n)$ is discrete uniform. (*cf.* § 7·8).

Example 7·52. *Suppose X is a non-negative integral valued random variable. Show that the distribution of X is geometric if it 'lacks memory', i.e., if for each* $k \geq 0$ *and* $Y = X - k$ *one has*

$$P(Y = t \mid X \geq k) = P(X = t), \text{ for } t \geq 0$$

[Madras Univ. B.Sc. (Main), 1988]

Solution. Let us suppose

$$P(X = r) = p_r;\ r = 0, 1, 2, \ldots$$

Define

$$q_k = P(X \geq k) = p_k + p_{k+1} + \ldots \qquad \ldots(*)$$

We are given

$$P(Y = t \mid X \geq k) = P(X = t) = p_t \qquad \ldots(**)$$

We have

$$P(Y = t \mid X \geq k) = \frac{P(Y = t \cap X \geq k)}{P(X \geq k)} = \frac{P(X - k = t \cap X \geq k)}{P(X \geq k)}$$

$$= \frac{P(X = k + t)}{P(X \geq k)} = \frac{p_{k+t}}{q_k}$$

$$\Rightarrow \qquad p_t = \frac{p_{k+t}}{q_k}, \qquad \text{[Using (**)]}$$

for every $t \geq 0$ and all $k \geq 0$. In particular, taking $k = 1$, we get

$$p_{t+1} = q_1 \cdot p_t = (p_1 + p_2 + \ldots) p_t = (1 - p_0) p_t \qquad \text{[From (*)]}$$

$$\Rightarrow \qquad p_t = (1 - p_0) p_{t-1} = (1 - p_0)^2 p_{t-2} = \ldots = (1 - p_0)^t p_0$$

Hence $\qquad p_t = P(X = t) = p_0 (1 - p_0)^t;\ t = 0, 1, 2, \ldots$

$\Rightarrow \qquad X$ has a geometric distribution.

EXERCISE 7 (d)

1. (*a*) If the probability that a target is destroyed on any one shot is 0·5, what is the probaility that it would be destroyed on 6th attempt?

Ans. $(0·5)^6$

(*b*) A couple decides to have children until they have a male child. What is the probability distribution of the number of children they would have? If the probability of a male child in their community is 1/3, how many children are they expected to have before the first male child is born?

(Sardar Patel U.B.Sc. Nov. 1991)

(*c*) Let X be a discrete random variable having geometric distribution with parameter p. Obtain its mean and variance. Also, show that for any two positive integers s and t,

$$P[X > s + t | X > s] = P[X > t]$$

2. The following distribution relates to the number of accidents to 650 women working on highly explosive shells during 5-week period. Show that a negative binomial distribution, rather than a geometric distribution, gives a very good fit to the data. How would you explaint this ?

Number of accidents :	0	1	2	3	4	5
Frequency :	450	132	41	22	3	2

(South Gujarat Univ., B.Sc. 1991)

3. (*a*) Show that the mean and variance of the geometric distribution

$$p(x) = q^x p \,;\; x = 0\,1, 2, \ldots$$

are respectively qp^{-1}, qp^{-2} (Allahabad Univ. B.Sc., 1989)

(*b*) Show that the mode of the distribution.

$$p(x) = (\tfrac{1}{2})^x \,; x = 1, 2, 3, \ldots$$

is 1.

4. Find (*i*) the probability generating function, (*ii*) the moment generating function, and (*iii*) the cumulant generating function for discrete random variable X following the geometric distribution

$$P(X = r) = (1 - p)\, p^{r-1} \,;\; r = 1, 2, \ldots$$

5. X_1 and X_2 are independent random variables with the same distribution $q^k p \,; k = 0, 1 \ldots$ Let Y be defined as the largest of X_1 and X_2, *i.e.*, $Y = \max (X_1, X_2)$. Obtain the joint distribution of Y and X_1 and the distribution of Y.

6. Identify the distributions with the following M.G.F.

$$e^t (5 - 4e^t)^{-1}.$$

Ans. Geometric Distribution, $p = 1/5$,

7. Prove the recurrence formula for Geometric Distribution, *viz.*,

$$p(x + 1) = q.p(x)$$

Let X and Y be independent random variables such that

$$P(X = r) = P(Y = r) = q^r p \,;\; r = 0, 1, 2, \ldots$$

p and q are positive numbers such that $p + q = 1$. Find (*i*) the distribution of $X + Y$ and (*ii*) the conditional distribution of X given $X + Y = 3$.

9. A die is cast until 6 appears. What is the probability that it must be cast more than five times.

$$\text{Ans. } P(X > 5) = 1 - P(X \le 5) = 1 - \sum_{x=1}^{5} (5/6)^{x-1} \cdot (1/6)$$

10. For the geometric distribution with p.m.f.

$$f(x) = 2^{-x} ; \; x = 1, 2, 3, \ldots$$

show that Chebychev's inequality gives

$$P(|X - 2| \leq 2) > \tfrac{1}{2}$$

while the actual probability is 15/16.

[Rajasthan Univ. B.Sc. (Hons.) 1992]

11. The conditional distribution of random variable X given $Y = y$ is $\dfrac{e^{-y} y^x}{x!}$ and the marginal probability density of Y is e^{-y}, where X is a discrete variable, *i.e.*, $x = 0, 1, 2, \ldots$ and Y is continuous, $y \geq 0$.

Show that the marginal distribution of X is geometric.

Hint. $g(x, y) = f(x \mid y)\, h(y) = \dfrac{e^{-y} y^x}{x!} \cdot e^{-y}$

$$\therefore \quad f(x) = \int_0^\infty \frac{e^{-2y} y^x}{x!}\, dy = \frac{1}{x!} \int_0^\infty e^{-2y} y^x\, dy = \frac{1}{x!} \cdot \frac{x!}{2^{x+1}}$$

12. If X and Y be two independent random variables, each representing the number of failures preceding the first success in a sequence of Bernoulli trials with p as probability of success in a single trial and q as probability of failure, show that $P(X = Y) = \dfrac{p}{1 + q}$

[Delhi Univ. B.Sc. (Stat. Hons.) 1993, '87]

Hint. We have $P(X = r) = P(Y = r) = q^r \cdot p$; $(r = 0, 1, 2, \ldots)$

$$P(X = Y) = \sum_{r=0}^{\infty} P(X = r \cap Y = r) = \sum_{r=0}^{\infty} P(X = r) \cdot P(Y = r)$$

[$\because$ X and Y are independent *r.v.'s*]

$$= p^2 \sum_{r=0}^{\infty} q^{2r} = p^2 (1 + q^2 + q^4 + \ldots) = \frac{p^2}{1 - q^2} = \frac{p}{1 + q}.$$

7·6. Hypergeometric Distribution. When the population is finite and the sampling is done without replacement, so that the events are stochastically dependent, although random, we obtain hypergeometric distribution. Consider an urn with N balls, M of which are white and $N - M$ are red. Suppose that we draw a sample of n balls at random (*without replacement*) from the urn, then the probability of getting k white balls out of n, $(k < n)$ is

$$\binom{M}{k} \binom{N - M}{n - k} \div \binom{N}{n}.$$

Definition. *A discrete random variable X is said to follow the hypergeometric distribution if it assumes only non-negative values and its probability mass function is given by*

$$P(X=k)=h(k;N,M,n)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}};\ k=0,1,2,\ldots,\min(n,M).$$

$$=0,\text{ otherwise} \qquad \ldots(7{\cdot}28).$$

Remarks. 1. N, M and n are known as the three parameters of hypergeometric distribution.

2. As it can be shown that

$$\sum_{k=0}^{n}\binom{M}{k}\binom{N-M}{n-k}\div\binom{N}{n}=1\ ,$$

this assignment of probabilities is permissible.

7·6·1. Mean and Varaince of the Hypergeometric Distribution.

$$E(X)=\sum_{k=0}^{n}k\cdot P(X=k)=\sum_{k=0}^{n}k\left\{\binom{M}{k}\binom{N-M}{n-k}\div\binom{N}{n}\right\}$$

$$=\frac{M}{\binom{N}{n}}\sum_{k=1}^{n}\left\{\binom{M-1}{k-1}\binom{N-M}{n-k}\right\}$$

$$=\frac{M}{\binom{N}{n}}\sum_{x=0}^{m}\binom{A}{x}\binom{N-A-1}{m-x},$$

where $x=k-1$, $m=n-1$, $M-1=A$

$$=\frac{M}{\binom{N}{n}}\cdot\binom{N-1}{m}=\frac{M}{\binom{N}{n}}\binom{N-1}{n-1}=\frac{nM}{N}$$

$$E\{X(X-1)\}=\sum_{k=0}^{n}k(k-1)\left\{\binom{M}{k}\binom{N-M}{n-k}\div\binom{N}{n}\right\}$$

$$=\frac{M(M-1)}{\binom{N}{n}}\sum_{k=2}^{n}\left\{\binom{M-2}{k-2}\binom{N-M}{n-k}\right\}$$

$$=\frac{M(M-1)}{\binom{N}{n}};\binom{N-2}{n-2}=\frac{M(M-1)n(n-1)}{N(N-1)}$$

---

*Since k white balls can be drawn from 'M' white balls in $\binom{M}{k}$ ways and out of the remaining $N-M$ red balls, $(n-k)$ can be chosen in $\binom{N-M}{n-k}$ ways, the total number of favourable cases is $\binom{M}{k}\times\binom{N-M}{n-k}$.

$$\therefore \quad E(X^2) = E[X(X-1)] + E(X) = \frac{M(M-1)n(n-1)}{N(N-1)} + \frac{nM}{N}$$

Hence $$V(X) = \frac{M(M-1)n(n-1)}{N(N-1)} + \frac{nM}{N} - \left(\frac{nM}{N}\right)^2$$

$$= \frac{NM(N-M)(N-n)}{N^2(N-1)} \qquad \text{(On simplification)}$$

7·6·2. Factorial Moments of Hypergeometric Distribution. The rth factorial moment is

$$E[X^{(r)}] = \sum_{k=r}^{n} k^{(r)} P(X=k) = \sum_{k=r}^{n} k^{(r)} \left\{\binom{M}{k}\binom{N-M}{n-k} + \binom{N}{n}\right\}$$

$$= \sum_{k=r}^{n} M^{(r)} \left\{\binom{M-r}{k-r}\binom{N-M}{n-k} + \binom{N}{n}\right\}^{*}$$

$$= M^{(r)} \sum_{j=0}^{n-r} \left\{\binom{M-r}{j}\binom{(N-r)-(M-r)}{(n-r)-j} + \binom{N}{n}\right\}, \text{ where } j = k-r$$

$$= \frac{M^{(r)} n^{(r)}}{N^{(r)}} \sum_{j=0}^{n-r} \left\{\binom{M-r}{j}\binom{(N-r)-(M-r)}{(n-r)-j} + \binom{N-r}{n-r}^{**}\right\}$$

$$= \frac{M^{(r)} n^{(r)}}{N^{(r)}} \sum_{j=0}^{n-r} h(j; N-r, M-r, n-r) = \frac{M^{(r)} n^{(r)}}{N^{(r)}} \cdot 1$$

$$\therefore \qquad E[X^{(r)}] = \frac{M^{(r)} n^{(r)}}{N^{(r)}} \qquad \ldots(7\cdot28\,a)$$

$$\Rightarrow \qquad \mu_x = E(X) = \frac{nM}{N}$$

$$E[X^{(2)}] = \frac{M(M-1)n(n-1)}{N(N-1)}$$

$$\sigma_X^2 = E[X^{(2)}] + E(X) - [E(X)]^2 = n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \cdot \frac{N-n}{N-1}$$

(On simplification) $\ldots(7\cdot28\,b)$

Remark. If we sample the n balls with replacement and denote by Y the number of white balls in the sample, then Y is a binomial variate with parameters n and p where

---

* $\because \; k^{(r)}\binom{M}{k} = \dfrac{k(k-1)(k-2)\ldots(k-r+1)M!}{k!(M-k)!}$

$= \dfrac{M(M-1)(M-2)\ldots(M-r+1)(M-r)!}{(k-r)!(M-k)!} = M^{(r)}\binom{M-r}{k-r}$

** $n^{(r)}\binom{N}{n} = N^{(r)}\binom{N-r}{n-r}$

$$p = M/N,\ q = 1 - p = (N - M)/N$$

$$\therefore \qquad E(Y) = np = \frac{nM}{N} = E(X)$$

$$\sigma_Y^2 = npq = n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \geq \sigma_X^2, \qquad \text{[From (7·28 } b\text{)]}$$

equality holding only if $n = 1$.

7·6·3. Approximation to Binomial Distribution. Hypergeometric distribution tends to binomial distribution as $N \to \infty$ and $\dfrac{M}{N} \to p$.

$$\begin{aligned}
h(k; N, M, n) &= \binom{M}{k}\binom{N-M}{n-k} \div \binom{N}{n} \\
&= \frac{M!}{k!(M-k)!} \cdot \frac{(N-M)!}{(n-k)!(N-M-n+k)!} \cdot \frac{n!(N-n)!}{N!} \\
&= \frac{M(M-1)(M-2)\dots(M-k+1)}{k!} \\
&\quad \times \frac{(N-M)(N-M-1)\dots(N-M-n+k+1)}{(n-k)!} \\
&\quad \times \frac{n!}{N(N-1)(N-2)\dots(N-n+1)} \\
&= \frac{n!}{k!(n-k)!} \cdot \frac{M}{N}\left(\frac{M}{N} - \frac{1}{N}\right)\left(\frac{M}{N} - \frac{2}{N}\right)\dots\left(\frac{M}{N} - \frac{k-1}{N}\right) \\
&\quad \times \frac{\left(1 - \frac{M}{N}\right)\left(1 - \frac{M}{N} - \frac{1}{N}\right)\dots\left(1 - \frac{M}{N} - \frac{(n-k-1)}{N}\right)}{\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\dots\left(1 - \frac{n-1}{N}\right)}
\end{aligned}$$

Proceeding to the limit as $N \to \infty$ and putting $\dfrac{M}{N} = p$, we get

$$\begin{aligned}
\lim_{N \to \infty} h(k; N, M, n) &= \binom{n}{k} \underbrace{p \cdot p \dots p}_{k \text{ times}} \underbrace{(1-p)(1-p)\dots(1-p)}_{(n-k) \text{ times}} \\
&= \binom{n}{k} p^k (1-p)^{n-k} = b(k; p, 1-p)
\end{aligned}$$

7·6·4. Recurrence Relation for the Hypergeometric Distribution. We have

$$h(k; N, M, n) = \binom{M}{k}\binom{N-M}{n-k} \div \binom{N}{n}$$

$$h(k+1; N, M, n) = \binom{M}{k+1}\binom{N-M}{n-k-1} \div \binom{N}{n}$$

$$\therefore \quad \frac{h(k+1; N, M, n)}{h(k; N, M, n)} = \frac{(n-k)(M-k)}{(k+1)(N-M-n+k+1)},$$

which is the required recurrence relation.

Example 7·53. *Explain how you will use hypergeometric model to estimate the number of fish in a lake.*

Solution. Let us suppose that in a lake there are N fish, N unknown. The problem is to estimate N. A catch of 'r' fish (all at the same time) is made and these fish are returned alive into the lake after marking each with a red spot. After a reasonable period of time, during which these marked' fish are assumed to have distributed themselves 'at random' in the lake, another catch of 's' fish (again, all at once) is made. Here r and s are regarded as fixed predetermined constants. Among these s fish caught, there will be, (say), X marked fish where X is a random variable following discrete probability function given by hypergeometric model:

$$f_X(x \mid N) = \binom{r}{x}\binom{N-r}{s-r} \div \binom{N}{s} = p(N), \text{ say} \qquad \ldots(*)$$

where x is an integer such that $\max(0, s-N+r) \le x \le \min(r, s)$ and $f_X(x \mid N) = 0$ otherwise.

The value of N is estimated by the principle of Maximum Likelihood (*c.f.* Chapter 15), *i.e.*, we find the value $\hat{N} = \hat{N}(x)$ of N which maximises $p(N)$. Since N is a discrete *r.v.*, the principle of maxima and minima in calculus cannot be used here. Here we proceed as follows :

$$\lambda(N) = \frac{p(N)}{p(N-1)} = \frac{(N-r)(N-s)}{N(N-r-s+x)}. \qquad \text{(On simplification)}$$

$$\therefore \quad \lambda(N) > 1 \text{ iff } N > \frac{rs}{x} \Rightarrow p(N) > p(N-1) \text{ iff } N > \frac{rs}{x} \qquad \ldots(i)$$

$$\text{and } \lambda(N) < 1 \text{ iff } N < \frac{rs}{x} \Rightarrow p(N) < p(N-1) \text{ iff } N < \frac{rs}{x} \qquad \ldots(ii)$$

From (*i*) and (*ii*) we see that $p(N) = f_X(x \mid N)$ reaches the maximum value (as a function of N) when N is approximately equal to rs/x. Hence maximum likelihood estimate of N is given by

$$\hat{N} = \frac{rs}{x} \Rightarrow \hat{N}(X) = \frac{rs}{X}$$

EXERCISE 7 (e)

1. (*a*) What is a hypergeometric distribution ? Find the mean and variance of this distribution. How is this distribution related to the binomial ?

[Nagarjuna Univ. M.Sc. 1991; Delhi Univ. B.Sc. (Stat. Hons.), 1989]

(*b*) Obtain binomial distribution as a limiting case of hyper-geometric distribution. [Delhi Univ. B.Sc. (Stat. Hons.), 1989,' 87]

2. Suppose that rockets of a certain type have, by many tests, been established as 90% reliable. Now a modification of the rocket design is being considered. Which of the following sets of evidence throws more doubt on the hypothesis that the modified rocket is only 90% reliable :

(*i*) Of 100 modified rockets tested, 96 performed satisfactorily.

(*ii*) Of 64 modified rockets tested, 62 performed satisfactorily ?

3. A taxi cab company has 12 Ambassadors and 8 Fiats. If 5 of these taxi cabs are in the shop for repairs and Ambassador is as likely to be in for repairs as a Fiat, what is the probbaility that·

(i) 3 of them are Ambassadors and 2 are Fiats ?

(ii) at least 3 of them are Ambassadors ? and

(iii) all 5 of them are of the same make ?

Ans. (i) $\binom{12}{3}\binom{8}{2}\div\binom{20}{5}$; (ii) $\sum_{x=3}^{5}\binom{12}{x}\binom{8}{5-x}\div\binom{20}{5}$

4. (a) Show how the hypergeometric disrtribution arises, by giving an example. Obtain the frequency function of a random variable X following the above law. Derive $E(X)$ and $V(X)$. Show that under certain conditions to be stated, the Binomial and Poisson distributions are special cases of the hypergeometric distribution. [Dibrugarh Univ. B.Sc. 1992]

(b) Find the factorial moments of the hypergeometric distribution.

[Delhi Univ. B.Sc. (Stat Hons.), 1993]

5. (a) Suppose that from a population of N elements of which M are defective and $(N-M)$ are non-defective, a sample of size n is drawn without replacement. What is the probability that the sample contains exactly x defectives? Name this probability distribution.

(b) Show that, for the distribution derived in (a),

$$E(X)=\frac{nM}{N} \text{ and } (ii)\ V(X)=\frac{nM}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n-1}{N-1}\right)$$

(c) Show that, under certain conditions to be stated, the binomial distribution may be looked upon as a limiting form of the probability distribution as derived in (a).

6. (a) 200 students of the F.Y. B.Sc. class in a certain College are divided at random into 20 batches of 10 each for the annual practical examination in Statistics. Suppose the class consists of 40 resident students and 160 non-resident students; and let R denote the number of resident students in the first batch. Use the binomial approximation to find the probability that $R \geq 3$.

Hint. The probability distribution of R is hyper-geometric with parameters : $N = 200$, $n = 10$, $M = 40$

Since $N\ (= 200)$ is large, the hypergeometric distribution (*) can be approximated by binomial distribution with parameters $n = 10$, $p = M/N = 40/200 = 0{\cdot}2$

$$\therefore \quad P(R=r)=\binom{10}{r}(0{\cdot}2)^r(0{\cdot}8)^{10-r};\ r=0,1,\ldots,10,$$

and required probability is :

$$P(R\geq 3)=1-[P(R=0)+P(R=1)+P(R=2)]=0{\cdot}323$$

(b) Find the probability that the income-tax official will catch 3 income-tax returns with illegitimate deductions, if he randomly selects 5 returns from among 12 returns of which 6 contain illegitimate deductions.

Ans. $\binom{6}{3}\binom{6}{2}\div\binom{12}{5} = 25/66.$

(c) If X and Y are independent binomial variates with parameters (n_1, p) and (n_2, p) respectively, find $P(X = r \mid X + Y = n)$

Ans. $\binom{n_1}{r}\binom{n_2}{n-r}\div\binom{n_1+n_2}{n}$

7. From a finite population of N animals in a given region, W are caught, marked and then released again. The animals are caught again one by one until m (pre-assigned) marked animals are caught. The total number of animals caught is a random variable X. Find P (X = n), for $m \le n \le N - W + m$

(Shivaji Univ. B.Sc., 1987)

Hint. $P(X = n) = P$ {Catching $(n-1)$ animals of whom $(m-1)$ are marked}$\times$ P {Catching the marked animal from the remaining $N - (n-1)$ animals}.

$$= \frac{\binom{W}{m-1}\binom{N-W}{n-m}}{\binom{N}{n-1}} \times \frac{\{W-(m-1)\}}{\{N-(n-1)\}}$$

$$= \binom{N-n}{W-m}\binom{n-1}{m-1}\div\binom{N}{W}$$

8. An urn contains M balls numbered 1 to M, where the first k balls are defective and the remaining $(M-K)$ are non-defective. A sample of n balls is drawn from the urn. Let A_k be the event that the sample of n balls contains exactly k defectives. Find $P(A_k)$ when the sample is drawn (i) with replacement and (ii) without replacement. [Delhi Univ. B.Sc. (Maths Hons.), 1989]

Hint. If sampling is done without replacement, we get hyper-geometric probability model.

$$P(A_k) = \binom{K}{k}\binom{M-K}{n-k}\div\binom{M}{n}$$

If sampling is done with replacement, then $X \sim B(n,\ p = K/M)$

$$\therefore\ P(A_k) = \binom{n}{k}(K/M)^k \cdot \left(1 - \frac{K}{M}\right)^{n-k} = \binom{n}{k}\frac{K^k (M-K)^{n-k}}{M^n}$$

9. X is a random variable distributed according to hyper-geometric law:

$$P(X = x) = h(x; n, a, b) = \frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}};\ x = 0, 1, 2, \ldots$$

Obtain the recurrence formula :

$$h(x+1; n, a, b) = \frac{(n-x)(a-x)}{(x+1)(b-n+x+1)}\,h(x; n, a, b)$$

10. For the hypergeometric distribution

$$h(N\,;n,p,x)=\frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}\;;x=0,1,2,\ldots$$

Prove that $\mu_1'=np$ and $\mu_2=\dfrac{n(N-n)\,pq}{N-1}$.

11. Explain how you will use hypergeometric model to estimate the number of wild animals in a dense forest.

12. A box contains N items of which 'a' items are defective and 'b' are non-defective, $(a+b=N)$. A sample of n items is drawn at random. Let X be number of defective items in the sample. Obtain the probability distribution of X and obtain the mean of the distribution.

7·7. Multinomial Distribution. This distribution c · · regarded as a generalisation of Binomial distribution.

When there are more than two mutually exclusive out trial, the observations lead to multinomial distribution. Suppose $\ldots_k$ are k mutually exclusive and exhaustive outcomes of a trial with ive probabilities $p_1, p_2, \ldots, p_k$.

The probability that E_1 occurs x_1 times, E_2 occurs x_2 ti ... and E_k, occurs x_k times in n independent observations, is given by

$$p(x_1, x_2, \ldots, x_k)=c\,p_1^{x_1}\,p_2^{x_2}\ldots p_k^{x_k}$$

where $\sum x_i=n$ and c is the number of permutation of the events $E_1, E_2, \ldots,$

To determine c, we have to find the number of permutations of n objects which x_1 are of one kind, x_2 of another kind, ..., and x_k of the kth kind, which is given by

$$c=\frac{n\,!}{x_1\,!\,x_2\,!\ldots x_k\,!}$$

Hence

$$p(x_1, x_2, \ldots, x_k)=\frac{n!}{x_1\,!\,x_2\,!\ldots x_k\,!}\cdot p_1^{x_1}\,p_2^{x_2}\ldots p_k^{x_k}\ ,\ 0\le x_i\le n$$

$$=\frac{n\,!}{\prod\limits_{i=1}^{k}x_i\,!}\cdot\prod_{i=1}^{k}p_i^{x_i}\,,\qquad \sum_{i=1}^{k}x_i=n \qquad \ldots(7{\cdot}29)$$

which is the required probability function of the multinomial distribution. It is so-called since (7·29) is the general term in the multinomial expansion

$$(p_1+p_2+\ldots+p_k)^n,\ \sum_{i=1}^{k}p_i=1$$

Since, total probability is 1, we have

$$\sum_x p(x)=\sum_x\left[\frac{n!}{x_1\,!\,x_2\,!\ldots x_k\,!}p_1^{x_1}\,p_2^{x_2}\ldots p_k^{x_k}\right]=(p_1+p_2\ldots+p_k)^n=1. \qquad \ldots[7{\cdot}29(a)]$$

7·7·1. Moments of Multinomial Distribution. The moment generating function is given by

$$M_X(t) = M_{X_1, X_2, \ldots, X_k}(t_1, t_2, \ldots, t_k) = E\left[\exp\left\{\sum_{i=1}^{k} t_i X_i\right\}\right]$$

$$= \sum_{x}\left[\frac{n!}{x_1!\, x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \exp\left(\sum_{i=1}^{k} t_i x_i\right)\right]$$

$$= \sum_{x}\left[\frac{n!}{x_1!\, x_2! \ldots x_k!} (p_1 e^{t_1})^{x_1} \ldots (p_k e^{t_k})^{x_k}\right]$$

$$= (p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k})^n \qquad \ldots(7·30)$$

where $x = (x_1, x_2, \ldots, x_k)$ [On using (7·29 (a)],

Now $M_{X_1}(t_1) = M_X(t_1, 0, 0, \ldots, 0) = \left(p_1 e^{t_1} + p_2 + p_3 + \ldots + p_k\right)^n$

$$= \left[(1 - p_1) + p_1 e^{t_1}\right]^n \qquad \left(\because \sum_i p_i = 1\right)$$

$\Rightarrow X_1 \sim B(n, p_1)$ [By uniqueness theorem of m.g.f.]

Similarly, we shall get:

$X_i \sim B(n, p_i)$; $i = 1, 2, \ldots, k.$

$\Rightarrow \quad E(X_i) = np_i$ and Var $X_i = np_i(1 - p_i)$; $i = 1, 2, \ldots, k$

$$E(X_i X_j) = \left[\frac{\partial^2 M}{\partial t_i\, \partial t_j}\right]_{t=0}, \; i \neq j$$

$$= \left[np_i e^{t_i} (n - 1)(p_1 e^{t_1} + \ldots + p_k e^{t_k})^{n-2} p_j e^{t_j}\right]_{t=0}$$

$$= n(n - 1) p_i p_j$$

$$\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i) E(X_j) = n(n-1) p_i p_j - n^2 p_i p_j = -np_i p_j$$

$$\therefore \rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i}\, \sigma_{X_j}} = \frac{-np_i p_j}{\sqrt{np_i(1 - p_i)\, np_j(1 - p_j)}}$$

$$= -\left[\frac{p_i p_j}{(1 - p_i)(1 - p_j)}\right]^{1/2}$$

Example 7·54. *The trinomial distribution of two r.v.'s X* and *Y is given by :*

$$f_{X,Y}(x, y) = \frac{n!}{x!\, y!\, (n - x - y)!} p^x q^y (1 - p - q)^{n-x-y}$$

for $x, y = 0, 1, 2, \ldots, n$ *and* $x + y \leq n$,

where $0 \leq p$, $0 \leq q$ and $p + q \leq 1$.

(*i*) *Find the marginal distributions of X* and *Y*

(*ii*) *Find the conditional distributions of X* and *Y* *and obtain* $E(Y | X = x)$ *and* $E(X | Y = y)$

(*iii*) *Find the correlation coefficient between X* and *Y*.

[Delhi Univ. B.Sc. (Maths Hons.) 1988; Spl Course-Statistics 1989;' 85]

Solution. The joint m.g.f. of X and Y is given by :

$$M_{X,Y}(t_1,t_2) = E\left(e^{t_1X+t_2Y}\right) = \sum_{x=0}^{n}\sum_{y=0}^{n-x}(pe^{t_1})^x(qe^{t_2})^y(1-p-q)^{n-x-y}$$

$$= \left[pe^{t_1} + qe^{t_2} + (1-p-q)\right]^n \qquad ...(i)$$

$$M_X(t_1) = M(t_1, 0) = \{(1-p) + pe^{t_1}\}^n \Rightarrow X \sim B(n,p) \qquad ...(ii)$$

$$M_Y(t_2) = M(0, t_2) = \{(1-q) + qe^{t_2}\}^n \Rightarrow Y \sim B(n,q) \qquad ...(iii)$$

Observe that $M(t_1,t_2) \neq M(t_1,0) \times M(0,t_2) \Rightarrow X$ and Y are not independent.

(ii) The conditional distribution of X given $Y = y$ is given by :

$$f(X \mid Y=y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_{XY}(x,y)}{{}^nC_y\, q^y(1-q)^{n-y}} \qquad [\because Y \sim B(n,q)]$$

$$= \frac{(n-y)!}{x!\,(n-y-x)!}\left(\frac{p}{1-q}\right)^x\left(\frac{1-p-q}{1-q}\right)^{n-y-x}$$

$$= \binom{n-y}{x}\left(\frac{p}{1-q}\right)^x\left(1-\frac{p}{1-q}\right)^{n-y-x} \; ; x = 0, 1, ..., n$$

$$\Rightarrow \quad X \mid (Y=y) \sim B\left(n-y,\, p/(1-q)\right) \qquad ...(iv)$$

$$\Rightarrow E(X \mid (Y=y)) = (n-y)\cdot p/(1-q) \qquad ...(v)$$

Similarly, we shall get

$$f(Y \mid X=x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f(x,y)}{{}^nC_x\, p^x(1-p)^{n-x}} \qquad [\because X \sim B(n,p)]$$

$$= \binom{n-x}{y}\left(\frac{q}{1-p}\right)^y\left(1-\frac{q}{1-p}\right)^{n-x-y} \; ; y = 0, 1, .., n.$$

$$\Rightarrow \quad Y \mid (X=x) \sim B\left(n-x,\, q/(1-p)\right) \qquad ...(vi)$$

$$\Rightarrow E[Y \mid (X=x)] = (n-x)\, q/(1-p) \qquad ...(vii)$$

(iii) Correlation Coefficient ρ_{XY} :

Since $X \sim B(n,p)$, $E(X) = np$, $Var\ X = np(1-p)$

$Y \sim B(n,q)$, $E(Y) = nq$, $Var\ Y = nq(1-q)$

$$E(XY) = \left.\frac{\partial^2 M(t_1,t_2)}{\partial t_1\, \partial t_2}\right|_{t_1=t_2=0} = n(n-1)pq$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) = n(n-1)pq - n^2pq = -npq$$

$$\therefore \quad \rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{-npq}{\sqrt{np(1-p)nq(1-q)}} = -\left[\frac{pq}{(1-p)(1-q)}\right]^{1/2}$$

Note. Here $p + q \neq 1$.

Example 7·55. *If $X_1, X_2, ..., X_k$ are k independent Poisson variates with prameters $\lambda_1, \lambda_2, \ldots, \lambda_k$ respectively, prove that the conditional distribution $P(X_1 \cap X_2 \cap \ldots \cap X_k \mid X)$, where $X = X_1 + X_2 + \ldots + X_k$ is fixed, is multinomial.* **[Lucknow U. B.Sc. (Hons.), 1992]**

Solution. $P[X_1 \cap X_2 \cap \ldots \cap X_k \mid X = n]$

$$= P[X_1 = r_1 \cap X_2 = r_2 \cap \ldots \cap X_k = r_k \mid X = n]$$

$$= \frac{P[X_1 = r_1 \cap X_2 = r_2 \cap \ldots \cap X_k = r_k \cap X = n]}{P(X = n)}$$

$$= \frac{P[X_1 = r_1 \cap \ldots \cap X_{k-1} = r_{k-1} \cap X_k = n - r_1 - r_2 \ldots - r_{k-1}]}{P(X = n)}$$

$$= \frac{P(X_1 = r_1)P(X_2 = r_2)\ldots P(X_{k-1} = r_{k-1})P(X_k = n - r_1 - \ldots - r_{k-1})}{P(X = n)}$$

($\because X_1, X_2, ..., X_k$ are independent)

Further, since $X_i, (i = 1, 2, ..., k)$ are independent Poisson variates with parameters λ_i respectively, $X = X_1 + X_2 + \ldots + X_k$ is also a Poisson variate with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_k = \lambda$ (say).

Hence $P[X_1 \cap X_2 \cap \ldots \cap X_k \mid X = n]$

$$= \frac{\dfrac{e^{-\lambda_1}\lambda_1^{r_1}}{r_1!} \cdots \dfrac{e^{-\lambda_{k-1}}\lambda_{k-1}^{r_{k-1}}}{r_{k-1}!} \cdot \dfrac{e^{-\lambda_k}\lambda_k^{n-r_1-\ldots-r_{k-1}}}{(n-r_1-\ldots-r_{k-1})!}}{\dfrac{e^{-\lambda}\lambda^n}{n!}}$$

$$= \left[\frac{n!}{r_1!\,r_2!\ldots r_{k-1}!\,(n-r_1-\ldots-r_{k-1})!}\right] \times \left[\left(\frac{\lambda_1}{\lambda}\right)^{r_1} \cdots \left(\frac{\lambda_{k-1}}{\lambda}\right)^{r_{k-1}} \left(\frac{\lambda_k}{\lambda}\right)^{n-r_1-\ldots-r_{k-1}}\right]$$

$$= \frac{n!}{r_1!\,r_2!\ldots r_k!} p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k}$$

where $\sum_{i=1}^{k} r_i = n$ and $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} \left(\frac{\lambda_i}{\lambda}\right) = \frac{1}{\lambda} \sum_{i=1}^{k} \lambda_i = 1$

Thus the conditional distribution $P(X_1 \cap X_2 \cap \ldots \cap X_k \mid X = n$ multinomial with probabilities $p_i = (\lambda_i/\lambda)$; $i = 1, 2, ..., k$ in k classes.

Remark. If X_i's are identically distributed independent Poisson variates with parameter m (say), then $\lambda_i = m$; $i = 1, 2, ..., k$ and $\lambda = \sum_{i=1}^{k} \lambda_i = km$.

$$\therefore \qquad p_i = \frac{\lambda_i}{\lambda} = \frac{1}{k}$$

Hence in this case the conditional distribution of $X_1, X_2, ..., X_k$, given that their sum $X_1 + X_2 + ... + X_k = n$, is a multinomial distribution with index n and the probability in each class being equal to $1/k$.

EXERCISE 7(f)

1. If $X_1, X_2, ..., X_k$ have a multinomial distribution with the parameters n and p_i $(i = 1, 2, ..., k)$ with $\Sigma p_i = 1$, obtain the joint probability

$$P(X_1 = x_1 \cap X_2 = x_2 \cap ... \cap X_k = x_k)$$

Obtain the corresponding moment generating function. Hence, or otherwise show that $E(X_i) = np_i$, $V(X_i) = np_i(1 - p_i)$

and $$\text{Cov}(X_i, X_j) = -np_i p_j, \ (i \neq j).$$

2. Discuss the marginal and conditional distributions, associated with the multinomial distribution. If $(n_1, n_2, ..., n_k)$ have a multinomial distribution with parameters $(n, p_1, p_2, ..., p_k)$ and if $c_i, d_i, i = 1, 2, ... k$ are constants, find the variance of $\sum_{i=1}^{k} c_i n_i$ and co-variance between $\sum_{i=1}^{k} c_i n_i$ and $\sum_{i=1}^{k} d_i n_i$.

3. If the random variables $X_1, X_2, ..., X_k$ have a multinomial distribution, show that the marginal distribution of X_i is a binomial distribution with the parameters n and p_i, with $i = 1, 2, ..., k$.

4. For the trinomial distribution of two r.v.'s X and Y given by:

$$f(x, y) = \frac{n!}{x!\,y!\,(n-x-y)!}\, p_1^x\, p_2^y\, p_3^{n-x-y}$$

where x and y are non-negative integers with $x + y \leq n$ and p_1, p_2, p_3 are proper positive fractions with $p_1 + p_2 + p_3 = 1$, and

$f(x, y) = 0$, otherwise

Show that (*i*) $X \sim B(n, p_1)$ and $Y \sim B(n, p_2)$

(*ii*) $X \mid (Y = y) \sim B(n - y, p_1/(1 - p_2))$ and $Y \mid (X = x) \sim B(n - x, p_2/(1 - p_1))$

(*iii*) $\rho(X, Y) = -\left[p_1 p_2/(1 - p_1)(1 - p_2)\right]^{1/2}$

4. If 'n' dice, each of which has 6 faces marked 1 to 6 are thrown, find the probability of getting a sum 's' on them.

Hint. The exhaustive number of ways in which n dice can fall is 6^n.

Since the total number of permutations in which six numbers, *viz.*, 1, 2, ..., 6 taken 'n' at a time can add to s is the coefficient of x^s in the multinomial expansion of $(x + x^2 + ... + x^6)^n$, the required number of favourable cases for getting a sum 's' on a dice is the co-efficient of x^s in the expansion of $(x + x^2 + \dots + x^6)^n$.

$\therefore$ Required Probability $= \dfrac{1}{6^n}$ [co-efficient of x^s in $(x + x^2 + \dots x^6)^n$]

Now identically, we have :

$$x + x^2 + \dots + x^6 = x\,(1 + x + \dots + x^5) = x\left(\frac{1 - x^6}{1 - x}\right)$$

and by binomial expansion

$$x^n\,(1 - x^6)^n = \sum_{k=0}^{n} (-1)^k \,.\, {}^nC_k\, x^{n+6k}\ .$$

and $$(1-x)^{-n} = \sum_{r=0}^{\infty} {}^{-n}C_r(-x)^r = \sum_{r=0}^{\infty} {}^{n+r-1}C_r \,.\, x^r = \sum_{r=0}^{\infty} {}^{n+r-1}C_{n-1} \,.\, x^r$$

$$\therefore\quad \left[x\left(\frac{1-x^6}{1-x}\right)\right]^n = \sum_{k=0}^{n}\sum_{r=0}^{\infty} (-1)^k \,.\, {}^nC_k \,.\, {}^{n+r-1}C_{n-1} \,.\, x^{n+6k+r}$$

To find the co-efficient of x^s, we put $n + 6k + r = s$ *i.e.*, $n + r = s - 6k$

Thus the co-efficient of x^s in $(x + x^2 + \dots + x^6)^n$

$$= \sum_{k=0}^{(s-n)/6} (-1)^k \,.\, {}^nC_k \,.\, {}^{s-6k-1}C_{n-1},$$

summation being extended over the integral values of k not exceeding $(s - n)/6$.

Hence required probability $= \displaystyle\sum_{k=0}^{(s-n)/6} (-1)^k \,.\, {}^nC_k\, {}^{s-6k-1}C_r/6^n$

Remarks. 1. The probability of getting a sum 's' with a throw of n dice, each having 'f' faces marked 1 to f is the co-efficient of x^s in

$$\frac{1}{f^n}\left[\left(x + x^2 + \dots + x^f\right)^n\right]$$

2. If n dice have faces $f_1, f_2, \dots, f_n$ respectively, then the required probability of getting a sum 's' is the co-efficient of x^s in

$$\frac{1}{f_1, f_2 \dots f_n}\left[(x + x^2 + \dots + x^{f_1})(x + x^2 + \dots + x^{f_2}) \dots (x + x^2 + \dots + x^{f_n})\right]$$

6. What is the probability of obtaining a sum of 15 points by throwing five dice toegther?

Hint. The number of exhaustive cases in throwing of 5 dice is 6^5.

The number of ways in which the 5 dice thrown will give 15 points is the co-efficient of x^{15} in the expansion of $(x^1 + x^2 + x^3 + \dots + x^6)^5$.

Favourable number of cases

$\therefore \quad =$ coefficient of x^{15} in $(x + x^2 + \dots + x^6)^5$

$\quad\ =$ coefficient of x^{10} in $(1 + x + \dots + x^5)^5$

$$= \text{coefficient of } x^{10} \text{ in } (1-x^6)^5 (1-x)^{-5}$$

$$(1-x^6)^5 = (1 - {}^5C_1 x^6 + {}^5C_2 x^{12} - \ldots - x^{30}) = (1 - 5x^6 + 10x^{12} - \ldots - x^{30})$$

$$(1-x)^{-5} = 1 + 5x + \frac{5 \times 6}{2!} x^2 + \frac{5 \times 6 \times 7}{3!} x^3 + \frac{5 \times 6 \times 7 \times 8}{4!} x^4 + \ldots$$

$$+ \frac{5 \times 6 \times 7 \times \ldots \times 14}{10!} x^{10} + \ldots$$

$$= (1 + 5x + 15x^2 + 35x^3 + 70x^4 + \ldots + 1001x^{10} + \ldots)$$

$\therefore$ Favourable number of cases

$$= \text{co-efficient of } x^{10} \text{ in } (1 - 5x^6 + 10x^{12} - \ldots - x^{30})$$

$$\times (1 + 5x + \ldots + 70x^4 + \ldots + 1001x^{10} + \ldots)$$

$$= (1001 - 5 \times 70) = 651$$

Hence the required probability $= \dfrac{651}{6^5} = \dfrac{651}{7776}$

7. Four dice, each marked 1 to 6, are thrown together. Find the probability of a total count being

(*i*) Exactly 12 and

or (*ii*) More than or equal to 20.

8. Four tickets marked 00,01,10, 11 respectively are placed in a bag. A ticket is drawn at random five times, being replaced each time. Find the probability that the sum of the numbers on the tickets thus drawn is 23.

9. Show that the mode of the multinomial distribution is given by $x_1, x_2, \ldots x_k$, satisfying

$$np_i - 1 < x_i \le (n + k - 1) p_i ; \; i = 1, 2, \ldots, k$$

[In order to establish this, show that

$$p_i x_j \le p_j (x_i + 1) \text{ for } 1 \le i, j \le k]$$

7·8. Discrete Uniform Distribution. A random variable X is said to have uniform distribution on n points $\{x_1, x_2, \ldots, x_n\}$ if its p.m.f. is given by :

$$P(X = x_i) = \frac{1}{n}; \; i = 1, 2, \ldots, n \qquad \ldots(7·31)$$

For example, if X has a uniform distribution on the points $\{0, 1, 2, \ldots, n\}$, then $P(X = i) = \dfrac{1}{n+1}; \; i = 0, 1, 2, \ldots, n.$ $\qquad \ldots(7·31\,a)$

Such distributions can be conceived in practice if under the given experimental conditions, the different values of the random variable become euqally likely. Thus for a die experiment, and for an experiment with a deck of cards such distribution is appropriate.

7·9. Power Series Distribution. A discrete r.v. X is said to follow a *generalized power series distribution* (g.p.s.d.), if its probability mass function is given by—

$$P(X=x)=\begin{cases}\dfrac{a_x\,\theta^x}{f(\theta)} & ;\ x=0,1,2,\ldots;\ a_x\geq 0\\[2ex] 0, \text{ elsewhere}\end{cases} \qquad \ldots(7\cdot32)$$

where $f(\theta)$ is a generating function, *i.e.*,

$$f(\theta)=\sum_{x\in S} a_x\,\theta^x,\ \theta\geq 0 \qquad \ldots(7\cdot32a)$$

so that $f(\theta)$ is positive, finite and differentiable and S is a non-empty countable sub-set of non-negative integers.

Remarks 1. By taking proper choice of S and $f(\theta)$, the g.p.s.d. can be reduced to binomial, Poisson and logarithmic series distribution and their truncated forms.

2. An inflated powerseries distribution (p.s.d.), inflated at zero is given by

$$P(X=x)=\begin{cases}1-\alpha+\dfrac{\alpha\,a_0}{f(\theta)}, & x=0\\[2ex] \alpha\,\dfrac{a_x\,\theta^x}{f(\theta)}; & x=1,2,\ldots\end{cases} \qquad \ldots(7\cdot33)$$

where α $(0<\alpha\leq 1)$, is the inflation parameter.

3. The truncated p.s.d. is given by:

$$P(X=x\mid S)=\frac{a_x\,\theta^x}{f(\theta)}\bigg/ f(S),\ x\in S$$
$$=0,\ \text{otherwise}$$

$$\Rightarrow\quad P(X=x\mid S)=\frac{a_x\,\theta^x}{f_1(\theta)};\ x\in S,\ \text{where } f_1(\theta)=\sum_{x\in S} a_x\,\theta^x$$
$$=0,\ \text{otherwise} \qquad \ldots(7\cdot34)$$

7·9·1. Moment Generating Function of p.s.d.

$$M_X(t)=\sum_{x=0}^{\infty} e^{tx}\,P(X=x)=\sum_{x=0}^{\infty} e^{tx}\left\{a_x\,\theta^x/f(\theta)\right\}$$

$$=\frac{1}{f(\theta)}\sum_{x=0}^{\infty} a_x\,(\theta e^t)^x=\frac{f(\theta e^t)}{f(\theta)} \qquad \ldots(7\cdot35)$$

7·9·2. Recurrence Relation for Cumulants of p.s.d. The cumulant generating function is given by

$$K_X(t)=\log M_X(t)=\log\left[\frac{f(\theta e^t)}{f(\theta)}\right]$$

$$\sum_{r=1}^{\infty}\kappa_r\,\frac{t^r}{r!}=\log f(\theta e^t)-\log f(\theta) \qquad \ldots(1)$$

Differentiating (1) partially w.r.to θ and t respectively, we get

$$\sum_{r=1}^{\infty} \frac{\partial}{\partial \theta} \kappa_r \frac{t^r}{r!} = \frac{e^t f'(\theta e^t)}{f(\theta e^t)} - \frac{f'(\theta)}{f(\theta)} \quad ...(2)$$

and
$$\sum_{r=1}^{\infty} \kappa_r \frac{r\, t^{r-1}}{r!} = \frac{\theta e^t f'(\theta e^t)}{f(\theta e^t)} \quad ....(3)$$

Subtracting (3) from θ times (2), we get

$$\theta \sum_{r=1}^{\infty} \frac{\partial}{\partial \theta} \kappa_r \frac{t^r}{r!} = \sum_{r=1}^{\infty} \kappa_r \cdot \frac{t^{r-1}}{(r-1)!} - \frac{\theta \cdot f'(\theta)}{f(\theta)}$$

Comparing like powers of t on both sides, we get

$$0 = \kappa_1 - \frac{\theta f'(\theta)}{f(\theta)} \Rightarrow \kappa_1 = \frac{\theta f'(\theta)}{f(\theta)} \quad ...(7·36)$$

and $\kappa_{r+1} = \theta \cdot \dfrac{d}{d\theta} \kappa_r \,;\ r = 1, 2, \ldots$ (Comparing co-efficient of $t^r/r\,!$)

Remark. We have

$$\text{Mean} = \kappa_1 = \frac{\theta f'(\theta)}{f(\theta)} \quad ...(7·36\,a)$$

Alternatively

$$\text{Mean} = \sum_{x=0}^{\infty} x \left\{ a_x\, \theta^x / f(\theta) \right\} = \frac{\theta}{f(\theta)} \sum_{x=0}^{\infty} x\, a_x\, \theta^{x-1} = \frac{\theta f'(\theta)}{f(\theta)}$$

7·9·3. Particular Cases of g.p.s.d. 1. *Binomial Distribution.* Let us take $\theta = p/(1-p)$, $f(\theta) = (1+\theta)^n$ and $S = \{0, 1, 2, \ldots, n\}$, a set of $(n+1)$ non – negative integers then

$$f(\theta) = \sum_{x \in S} a_x\, \theta^x \Rightarrow (1+\theta)^n = \sum_{x=0}^{n} a_x\, \theta^x$$

$$\Rightarrow \qquad a_x = \binom{n}{x}$$

$$\therefore\ P(X = x) = \frac{\binom{n}{x}\left[\dfrac{p}{(1-p)}\right]^x}{\left[1 + \dfrac{p}{(1-p)}\right]^n}$$

$$= \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}\,; & x = 0, 1, \ldots, n \\ 0, \text{ otherwise} \end{cases}$$

which is the probability mass function of the binomial distribution with parameters n and p.

2. Negative Binomial Distribution. Let us take $\theta = p/(1+p)$, $f(\theta) = (1-\theta)^{-n}$ and $S = \{0, 1, 2, \ldots \text{ ad infinity}\}$, $0 \le \theta < 1$, $n > 0$.

Now $$f(\theta) = \sum_{x \in S} a_x \theta^x \Rightarrow (1-\theta)^{-n} = \sum_{x=0}^{\infty} a_x \theta^x$$

$$\Rightarrow \quad a_x = (-1)^x \binom{-n}{x} = (-1)^x \cdot (-1)^x \binom{n+x-1}{x} = \binom{n+x-1}{x}$$

$$\therefore \quad P(X=x) = \sum_{x=0}^{\infty} \binom{n+x-1}{x} \left[(p/1+p)\right]^x \Big/ \left[1 - \{p/(1+p)\}\right]^{-n}$$

$$= \sum_{x=0}^{\infty} \binom{n+x-1}{x} p^x (1+p)^{-(n+x)}; \; x = 0, 1, 2, \ldots$$

$$= \sum_{x=0}^{\infty} \binom{-n}{x} (1+p)^{-(n+x)} (-p)^x; \; x = 0, 1, 2, \ldots$$

which is the probability mass function of the negative binomial distribution.

3. Logarithmic Series Distribution. Let $f(\theta) = -\log(1-\theta)$ and $S = \{1, 2, 3, \ldots\}$.

Then $$f(\theta) = \sum_{x \in S} a_x \theta^x \Rightarrow -\log(1-\theta) = \sum_{x=1}^{\infty} a_x \theta^x, \text{ i.e., } a_x = \frac{1}{x}$$

$$\therefore P(X=x) = \begin{cases} \dfrac{a_x \theta^x}{f(\theta)} = \dfrac{\theta^x}{x[-\log(1-\theta)]}; & = 1, 2, 3, \ldots \\ 0, & \text{otherwise} \end{cases}$$

4. Poisson Distribution. Let $f(\theta) = e^{\theta}$, $S = \{0, 1, 2, \ldots\}$. Then

$$f(\theta) = \sum_{x \in S} a_x \theta^x \Rightarrow e^{\theta} = \sum_{x=0}^{\infty} a_x \theta^x \text{ i.e., } a_x = \frac{1}{x!}$$

$$\therefore \quad P(X=x) = \frac{a_x \theta^x}{f(\theta)} = \frac{\theta^x}{x! \, e^{\theta}} = \frac{e^{-\theta} \theta^x}{x!}; \; x = 0, 1, 2, \ldots$$

which is the probability mass function of the Poisson distribution with parameter θ.

ADDITIONAL EXERCISES ON CHAPTER VII

1. Show that the necessary and sufficient conditions for two given numbers a, b to be respectively the mean and the variance of some binomial distribution are that $a > b > 0$ and $\dfrac{a^2}{a-b}$ is an integer.

Show further that when these conditions are satisfied, the binomial distribution is uniquely determined.

2. In a game of taking a chance, a contestant has to give correct answers to 4 out of 5 questions to win the contest. Questions are given with 3 answers each, out of which one is a correct answer. If a contestant answers the questions by selecting the answers at random, what is the probability that he will win the contest?

Ans. $10/3^5 = 0{\cdot}0412$

3. Suppose the automatic machines of a plant fail with probability q, the machine failure is independent from machine to machine and the plant stays in operation, if at least half of the machines run. Consider a two-machine plant and a four-machine plant. Show that the value of q for uninterrupted operations,

(*i*) when the value of q is same in both plants is $\frac{1}{2}$,

(*ii*) when a two-machine plant is perferred is $q > \frac{1}{2}$, and

(*iii*) when a four-machine plant is preferred is $q < \frac{1}{2}$

4. If $b(r\,; n, p) = {}^nC_r\, p^r \cdot q^{n-r}$ is the binomial probability in the usual notation and if $B(k\,; n, p) = \sum_{r=0}^{k} b(r\,; n, p)$, prove the following results for the "tails" of the binomial distribution.

(*i*) $1 - B(k-1\,; n, p) \le \dfrac{n}{k-np}\, b(k\,; n, p),\ k > np+1$

(*ii*) $B(k\,; n, p) \le \dfrac{n}{np-k}\, b(k\,; n, p),\ k < np$

(*iii*) $1 - B(k\,; n, p) = n \dbinom{n-1}{k} \displaystyle\int_0^p t^k (1-t)^{n-k-1}\, dt$

5. If a coin is tossed n times where n is very large even number, show that the probability of getting exactly $(\frac{1}{2}n - p)$ heads and $(\frac{1}{2}n + p)$ tails is approximately

$$\left(\frac{2}{\pi n}\right)^{\frac{1}{2}} e^{-2p^2/}$$

6. If $X \sim B(n, p)$, show that
$P(X \le 2) = P[X \ge (n-2)]$, if and only if $p = \frac{1}{2}$.

[Calcutta Univ. B.Sc. (Hons.), 1989]

7. If $X \sim B(n, p)$, show that X is symmetrically distributed about c if and only if $p = 1/2$ and $c = n/2$. (Madurai Univ. M.A., 1991)

8. If $X \sim B(n, p)$, and $Y = X^2$, find corr. (X, Y)

[Delhi Univ. (Stat Hons.) Spl. Course; 1989]

9. A and B have equal chances of winning a single game, A wants n games and B, $n+1$ games to win a match. Show that the odds in favour of A are $1+P$ to $1-P$, where $P = \dfrac{(2n)!}{n!\,n!\,2^{2n}}$

Hint. The probability that A wins at least n games is

$${}^{2n}C_n q^n p^n + {}^{2n}C_{n+1} q^{n-1} p^{n+1} + \ldots + {}^{2n}C_{2n} p^{2n}$$

Now ${}^{2n}C_0 + {}^{2n}C_1 + \ldots + {}^{2n}C_{n-1} + {}^{2n}C_n + {}^{2n}C_{n+1} + \ldots + {}^{2n}C_{2n} = 2^{2n}$

$\therefore\ {}^{2n}C_{n+1} + {}^{2n}C_{n+3} + \ldots + {}^{2n}C_{2n} = \frac{1}{2}[2^{2n} - {}^{2n}C_n]$

$\therefore$ Probability of A's win $= \dfrac{1}{2^{2n}}\left(\frac{1}{2}(2^{2n} - {}^{2n}C_n)\right) = \frac{1}{2}(1 - P)$

$\therefore$ Probability of A's losing $= 1 - \frac{1}{2}(1 - P) = \frac{1}{2}(1 + P)$

Hence the result.

10. (*a*) The chance of success in each Bernoulli trial is p. If p_k is the probability that there are even number of successes in k trials, prove that

$$p_k = p + p_{k-1}(1 - 2p) \qquad \ldots(*)$$

Deduce that $p_k = \dfrac{1}{2}\left[1 + (1 - 2p)^k\right]$

(*b*) Also obtain the probability generating function of (*) and hence obtain an explicit expression for p_k

(*c*) Obtain an expression for p_k directly without using (*a*) or (*b*).

11. A spider and a fly are situated at the corners (0, 0) and (n, n) of a rectangular grid. The spider walks only north or east, the fly only south or west. They take their steps simultaneously to an adjacent vertex of the grid. Show that, if the successive steps are independent and equally likely to go in each of the two possible directions, the probability that they will meet is $\dbinom{2n}{n}\left(\dfrac{1}{2}\right)^{2n}$

[Delhi Univ. B.Sc. (Statistics Hons.) Spl. Course, 1988]

12. For the binomial distribution, show that the probability that the number of successes in n trials should not exceed x is given by

$$\frac{\displaystyle\int_{p/q}^{\infty} \frac{y^x}{(1 + y)^{n+1}}\, dy}{\displaystyle\int_0^{\infty} \frac{y^x}{(1 + y)^{n+1}}\, dy}, \quad p + q = 1$$

where p is the probability of success.

13. Prove the identity

$$p^n + \binom{n}{1} p^{n-1} q + \binom{n}{2} p^{n-2} q^2 + \ldots + \binom{n}{k} p^{n-k} q^k = \sum_{x=0}^{k} \binom{n}{x} \left(p^{n-x} q^x\right)$$

$$= \frac{\displaystyle\int_0^p x^{n-k-1}(1 - x)^k\, dx}{\displaystyle\int_0^1 x^{n-k-1}(1 - x)^k\, dx}$$

Hint. For Questions 12 and 13, see Example 7·23.

14. Let X be a random variable whose probability function is $b(x\,;n,p)$. Let $Y = X/n$ be a new random variable. Show that the expected value of Y is p and the variance of Y is pq/n. If $p(y)$ is the probability function for Y, show that $p(y) = b(ny\,;n,p)$. What are the possible values that y can take on ?

15. Suppose that the number of telephone calls that an operator receives from 9·00 to 9·05 hours in a day follows a Poisson distribution with mean 3. Find the probability that (*i*) the operator will receive no calls in that time interval tomorrow, (*ii*) In the next three days the operator will receive a total of 1 call in that time interval.

Ans. (*i*) e^{-3} (*ii*) $3 \times (e^{-3})^2 (1 - e^{-3})$.

16. A large number of observations on a given solution which contained bacteria were made taking samples 1 ml. each, noting down the number of bacteria present in each sample. Assuming the Poisson distribution, and given that 10% samples contained no bacteria, find the average number of bacteria per ml.

Ans. $\log_e 10$ or 2·3026

17. The number of oil tankers, say N, arriving at a certain refinery each day has a Poisson distribution with parameter 2. Present port facilities can service three tankers a day. If more than three tankers arrive in a day, the tankers in excess of three must be sent to another port.

(*i*) On a given day, what is the probability of having to send tankers away?

(*ii*) How much must present facilities be increased to permit handling all tankers on approximately 90 per cent of days ?

(*iii*) What is the expected number of tankers arriving per day ?

(*iv*) What is the expected number of tankers serviced daily? and

(*v*) What is the expected number of tankers turned away daily?

Ans. (*i*) 0·145, (*ii*) 4, (*iii*) 2, (*iv*) 1·785 and (*v*) 0·215.

18. If X is any non-negative integer valued variate and a is any positive number, show that

$$P(X \geq a) \leq t^{-a} . E(t^X)\,;\ t > 1$$

Verify the inequality

$$P(X \geq 2\lambda) \leq (e/4)^{\lambda} \text{ when } X \sim P(\lambda).$$

19. If X is any non-negative integer valued and a is any positive number, show that

$$P(X \leq a) \leq t^{-a}\, E(t^X),\ 0 < t \leq 1.$$

Verify the inequality :

$$P\left(X \leq \frac{1}{2} m\right) \leq (2/e)^{m/2}, \text{ when } X \sim P(m)$$

20. Suppose (X, Y) have the joint p.m.f.

$$f(x,y) = \frac{e^{-(a+b)}\, a^x\, b^{y-x}}{x\,!\,(y-x)\,!}, x = 0, 1, 2, \ldots\,; y = x, x+1, \ldots$$

Show that the correlation coefficient between X and Y is $[a/(a+b)]^{1/2}$.

Also obtain the distribution of $Y-X$.

[Delhi Univ. (Spl. Course. Statistics Hons.), 1988]

Hint. $$M_{X,Y}(t_1, t_2) = \sum_{x=0}^{\infty} \sum_{y=x}^{\infty} \left[e^{t_1 x + t_2 y} \frac{e^{-(a+b)} \cdot a^x b^{y-x}}{x!(y-x)!} \right]$$

$$= e^{-(a+b)} \left[\sum_{x=0}^{\infty} \frac{\left(a e^{t_1} e^{t_2}\right)^x}{x!} \sum_{z=0}^{\infty} \frac{\left(b e^{t_2}\right)^z}{z!} \right]; (y-x=z)$$

$$= \exp\left[a e^{t_1+t_2} + b e^{t_2} - a - b\right] \quad ...(**)$$

$$M_X(t_1) = M(t_1, 0) = \exp\left[a(e^{t_1} - 1)\right] \Rightarrow X \sim P(a)$$

$$M_Y(t_1) = M(0, t_2) = \exp\left[(a+b)\{e^{t_2} - 1\}\right] \Rightarrow Y \sim P(a+b)$$

$$E(XY) = \frac{\partial^2 M(t_1, t_2)}{\partial t_1 \partial t_2}\bigg|_{t_1 = t_2 = 0} = a^2 + ab + a$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = a^2 + ab + a - a(a+b) = a$$

Distribution of Y – X. Taking $t_1 + t_2 = 0 \Rightarrow t_1 = -t_2$ in (**),

$$M(-t_2, t_2) = E(e^{-t_2 X + t_2 Y}) = E(e^{(Y-X)t_2}) = \exp\left[b(e^{t_2} - 1)\right]$$

$$\Rightarrow \quad Y - X \sim P(b)$$

21. If X is Negative Binomial variate with parameters (k and Q^{-1}), prove that

$$P(X \geq m) = \frac{1}{B(m,k)} \cdot \int_0^P \frac{x^{m-1} dx}{(1+x)^{k+m}}; \ Q - P = 1$$

Hint. $$P(X \geq m) = \sum_{r=m}^{\infty} \binom{-k}{r} Q^{-k-r}(-P)^r$$

$$= \sum_{r=m}^{\infty} \binom{k+r-1}{r} Q^{-k-r} P^r$$

$$\frac{d}{dP}[P(X \geq m)] = \sum_{r=m}^{\infty} (T_r - T_{r+1}); T_r = r \cdot \binom{k+r-1}{r} P^{r-1} Q^{-k-r}$$

$$= T_m = \frac{1}{B(m,k)} \cdot P^{m-1} Q^{-k-m}$$

Integrating, we get

$$P(X \geq m) = \frac{1}{B(m,k)} \int_0^P \frac{x^{m-1}\, dx}{(1+x)^{k+m}} \qquad (\because Q - P = 1)$$

(b) If X is N.B. (k, p), show that

$$P(X \geq m) = \frac{1}{B(m,k)} \cdot \int_p^1 y^{k-1} (1-y)^{m-1}\, dy$$

22. In a sequence of independent trials, the probability of a success on each trial is 'p'. By considering the outcome of the first trial, show that $G_r(t)$, the p.g.f. of the number of trials required to achieve the rth success, satisfies :

$$G_r(t) = pt\, G_{r-1}(t) + qt\, G_r(t)$$

and hence obtain $G_r(t)$. [Delhi Univ. (Spl. Course Statistics Hons.), 1987]

Ans. $G_r(t) = [pt/(1-qt)]^r$

23. Let X and Y be independent random variables with the same (geometric) distribution given by $P(X = k) = pq^k$; $k = 0, 1, 2, \ldots$

Let $Z = \max(X, Y)$

(i) Find the probability distribution of Z.

(ii) Find the joint probability distribution of X and Z.

(iii) Find the conditional probability distribution of X given $Z = l$, i.e., compute $P(X = k \mid Z = l)$ for all $k, l = 0, 1, 2, \ldots$

(iv) Find the conditional probability distribution of Z given $X = k$, i.e. compute $P(Z = l \mid X = k)$ for all k, $l = 0, 1, 2, \ldots$

Ans. (i) $P(Z = l) = pq^l \left[2 - q^l - q^{l+1}\right]$; $l = 0, 1, 2, \ldots$

(ii) $$P(Z = l \cap X = k) = \begin{cases} 0, \text{ if } l < k \\ pq^k(1 - q^{k+1}) \text{ if } l = k = 0, 1, 2, \ldots \\ p^2 q^{k+l} \text{ if } l > k = 0, 1, 2, \ldots \end{cases}$$

(iii) $$P(X = k \mid Z = l) = \begin{cases} 0 \text{ if } l < k \\ (1 - q^{k+1})/(2 - q^k - q^{k+1}) \text{ if } l = k = 0, 1, 2, \ldots \\ pq^k/(2 - q^l - q^{l+1}) \text{ if } l > k = 0, 1, 2, \ldots \end{cases}$$

(iv) $$P(Z = l \mid X = k) = \begin{cases} 0 \text{ if } l < k \\ 1 - q^{k+1} \text{ if } l = k = 0, 1, 2, \ldots \\ pq^l \text{ if } l > k = 0, 1, 2, \ldots \end{cases}$$

24. Suppose that $X_1, X_2, \ldots, X_n$ are mutually independent indicator random variables, with $P(X_i = 1) = p$, $0 < p < 1$. Show that for $1 \leq M \leq N$,

$$P\left(\sum_{i=1}^{M} X_i = k \,\middle|\, \sum_{i=1}^{N} X_i = n\right) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

25. Suppose one makes $(m + n)$ independent trials of an experiment whose probability of success at each trial is p. Let $q = (1 - p)$. Show that for any $k = 0, 1, 2, \ldots n$, the conditional probability that exactly $(m + k)$ trials will result in success, given that the first m trials result in success, is equal to $\binom{n}{k} p^k q^{n-k}$. Show further that the conditional probability that exactly $(m + k)$ trials will result in success, given that at least m trials result in success is equal to

$$\binom{m+n}{m+k}\left(\frac{p}{q}\right)^k \Bigg/ \sum_{r=0}^{n} \binom{m+n}{m+r}\left(\frac{p}{q}\right)^r$$

26. Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli variates with common parameter $p = P(X_1 = 1)$. Let $S_j = X_1 + X_2 + \ldots + X_j$ for $1 \le j \le n$. Show that $P(S_j = r \mid S_n = s)$ does not depend on p $(0 < p < 1)$ and takes the form of a hypergeometric probability for $1 \le j \le n$, $0 \le r \le s \le n$.

Hint. $S_n \sim B(n, p)$

$$P(S_j = r \cap S_n = s) = \binom{j}{r} p^r q^{j-r} \cdot \binom{n-j}{s-r} p^{s-r} q^{n-j-s+r}$$

$$P(S_j = r \mid S_n = s) = P(S_j = r \cap S_n = s)/P(S_n = s)$$

$$= \binom{j}{r}\binom{n-j}{s-r} + \binom{n}{s};\ 1 \le j \le n, 0 \le r \le s \le n$$

a result, which is independent of p.

27. An urn contains w white balls and b black balls. Balls are drawn one at a time from the urn, without replacement. Find the distribution of the number X of draws needed to obtain the k th black ball. Find also the factorial moments $E\left[X^{(r)}\right]$. **[Delhi Univ. B.Sc. Statistics Hons. (Spl. Course), 1989]**

Ans. $$P(X = x) = \frac{\binom{b}{k-1}\binom{w}{x-k}}{\binom{b+w}{x-1}} \times \left(\frac{b-k+1}{b+w-x+1}\right)$$

$$= \binom{x-1}{k-1}\binom{b+w-x}{b-k} + \binom{b+w}{b};\ (\text{On simplification})$$

For $E\left[X^{(r)}\right]$, proceed as in § 7·6·2.

$$E\left[X^{(r)}\right] = k^{(r)}\,(b + w + 1)^{(r)}/(b + 1)^{(r)}$$

28. The joint p.m.f. of two discrete r.v.'s X_1 and X_2 is:

$$p(x_1, x_2) = \binom{n_1}{x_1}\binom{n_2}{x_2 - x_1} p^{x_2}(1 - p)^{n_1 + n_2 - x_2}$$

with $x_1 \le x_2 \le n_2 + x_1$; $0 \le x_1 \le n_1$.

Find the marginal distributions of X_1 and X_2.

Ans. $X_1 \sim B\,(n_1, p')$ and $X_2 \sim B\,(n_1 + n_2, p)$

29. Two discrete random variables X and Y have the joint probability distribution :

$$p(x,y) = \frac{9\,!}{x\,!\,y\,!\,(9 - x - y)\,!}\left(\frac{1}{3}\right)^9, \text{ where}$$

$$0 \le x \le 9,\ 0 \le y \le 9 \text{ and } 0 \le (x + y) \le 9$$

(*i*) Show that the marginal distribution of X is binomial with parameters 9 and ⅓.

(*ii*) Show that the conditional distribution of Y given $X = 3$ is also binomial with parameters 6 and ½.

30. A Polya process is defined by the quantities :

$$P_k(t) = \left[\frac{\lambda t}{1 + b\lambda t}\right]^k \frac{1(1 + b)\ldots\{1 + (k-1)b)\}}{k\,!}\,P_0(t)$$

where $P_0(t) = (1 + b\lambda t)^{-1/b}$

and λ, b are parameters, t is a continuous variable and k may take zero or poisitive integral values only. Verify that the distribution satisfies the requirements for a probability distribution in K and find the expectation of K and its variance.

Hint. Let K be a random variable with the distribution,
$P(K = k) = P_k(t)\,;\ k = 0, 1, 2, \ldots, \infty.$

$$\sum_{k=0}^{\infty} P_k(t) = (1 + b\lambda t)^{-1/b} \sum_{k=0}^{\infty}\left[\frac{\lambda t}{1 + b\lambda t}\right]^k \times \frac{b^k\left[\frac{1}{b}\right]\left[\frac{1}{b}+1\right]\ldots\left[\frac{1}{b}+k-1\right]}{k\,!}$$

$$= (1 + b\lambda t)^{-1/b} \sum_{k=0}^{\infty}\left[\frac{\lambda t}{1 + b\lambda t}\right]^k b^k\binom{(1/b)+k-1}{k}$$

$$= (1 + b\lambda t)^{-1/b}\,\frac{1}{\left[1 - \frac{b\lambda t}{1 + b\lambda t}\right]^{1/b}}\ * = 1$$

$\therefore P_k(t)$ represents a probability distribution for every fixed b, λ and t.

M.G.F. of $K = E(e^{Ku}) = \sum_{k=0}^{\infty} e^{uk} P_k(t)$.

$$*\,(1 - x)^{-n} = \sum_{k=0}^{\infty}\binom{-n}{k}(-x)^k = \sum_{k=0}^{\infty}\binom{n+k-1}{k}x^k$$

$$= (1+b\lambda t)^{-1/b} \sum_{k=0}^{\infty} \left[e^{ku} \frac{(\lambda t)^k}{(1+b\lambda t)^k} \frac{b^k \left[\frac{1}{b}\right]\left[\frac{1}{b}+1\right]\cdots\left[\frac{1}{b}+k-1\right]}{k!} \right]$$

$$= (1+b\lambda t)^{-1/b} \sum_{k=0}^{\infty} \left[\frac{e^u b\lambda t}{1+b\lambda t}\right]^k \binom{(1/b)+k-1}{k}$$

$$= (1+b\lambda t)^{-1/b} \frac{1}{\left[1-\frac{e^u b\lambda t}{1+b\lambda t}\right]^{1/b}} = [1+b\lambda t - e^u b\lambda t]^{-1/b}$$

$= g(u)$, (say)

$$\therefore \quad g'(u) = \frac{1}{b}\left[1+b\lambda t - e^u b\lambda t\right]^{-(1/b)-1} \cdot (e^u b\lambda t)$$

$$E(K) = \mu_1' \text{ (about origin)} = \text{Mean} = [g'(u)]_{u=0}$$

$$= \left[\frac{1}{b}\right]\left[1+b\lambda t - b\lambda t\right]^{-(1/b)-1} \cdot (b\lambda t) = \lambda t$$

Similarly $\quad \mu_2' = [g''(u)]_{u=0} = (b+1)\lambda^2 t^2 + \lambda t$

$\therefore \quad$ Variance $= \mu'_2 - \mu_1'^2 = \lambda t\,(1+b\lambda t)$.

OBJECTIVE TYPE QUESTIONS

1. (*i*) Match the correct parts to make a valid statement :

(*a*)	Binomial distribution applies to	1.	rare events
(*b*)	Poisson distribution applies to	2.	repeated two alternatives.
(c)	The mean of a Hypergeometric distribution	3.	$\frac{1-6pq}{npq}$
(*d*)	The moment generating function of negative binomial distribution	4.	$n \cdot \frac{M}{n}\left(1-\frac{M}{n}\right)\left(\frac{N-n}{N-1}\right)$
(*e*)	The coefficient of kurtosis of a binomial distribution	5.	$(Q - pe^t)^{-r}$
(*f*)	The variance of geometric distribution	6.	$\frac{nM}{N}$
(*g*)	Variance of Hypergeometric distribution	7.	$\frac{q}{p^2}$

II. Under what conditions binomial distribution tends to (*i*) Poisson distribution, (*ii*) Normal distribution, (*iii*) Geometric distribution. Give practical examples (one each) where you would expect binomial, Poisson, negative binomial and geometric distribution.

III. State the relationship between :

(*i*) Mean and variance of Poisson distribution.

(*ii*) Mean and variance of negative binomial distribution.

(*iii*) Mean and variance of geometric distribution.

(*iv*) Poisson distribution and binomial distribution.

(*v*) Hypergeometric distribution and binomial distribution.

IV. Name the discrete distribution for which

(*i*) Mean and variance have the same value.

(*ii*) Mean is greater than the variance.

V. State which of the following statements are *True* and which are *False*. In case of the false statement, give the correct statement :

(*i*) Mean of binomial distribution is 3 and variance is 5.

(*ii*) Mean of Poisson distribution is 2 and variance is 3.

(*iii*) The sum of two independent Poisson variates is also a Poisson variate. The result holds for the difference also.

(*iv*) For a binomial distribution,

Mean = Mode = Median

(*v*) The Poisson distribution is a limiting case of binomial distribution when $n \to \infty$, $p \to 0$, $np \to m$.

(*vi*) Nearly all the distributions are particular cases of Poisson distribution.

(*vii*) The sum of two binomial variates is a binomial variate if the variables are independent and have the different probabilities of success.

(*viii*) Negative binomial distribution may be regarded as the generalisation of geometric distribution.

VI. Fill in the blanks :

(*i*) The variance of a binomial distribution is ....

(*ii*) The β-coefficient of skewness of the binomial distribution is ....

(*iii*) The moment generating function of Poisson distribution is ....

(*iv*) The characteristic function of negative binomial distribution is ....

(*v*) The coefficient of skewness of a Poisson distribution is ....

(*vi*) Poisson distribution is a limiting case of binomial distribution under the conditions ....

(*vii*) For Poisson distribution all cumulants ....

(*viii*) Mean > variance for .... distribution.

(*ix*) For the Poisson distribution, the variance and the third central moment are ....

(*x*) Mean < variance for .... distribution.

VII. Give the correct answer to each of the following :

(*i*) The skewness in a binomial distribution will be zero, if

(*a*) $p < \frac{1}{2}$, (*b*) $p = \frac{1}{2}$, (*c*) $p > \frac{1}{2}$, (*d*) $p < q$.

(*ii*) The mean and variance of negative binomial distribution :

(*a*) are same, (*b*) cannot be same, (*c*) are sometimes equal in limiting case, as $n \to \infty$.

(*iii*) The characteristic function of Poisson distribution $P(m)$ is

(*a*) $e^{m(it-1)}$, (*b*) $e^{m(e^{it}-1)}$, (*c*) e^{mit}, (*d*) none of these.

(*iv*) The coefficient of variation of Poisson distribution with mean 4 is

(*a*) $\frac{1}{4}$, (*b*) $\frac{2}{4}$, (*c*) 4, (*d*) 2

(*v*) The coefficient of kurtosis of a Poisson distribution with mean m is

(*a*) $1/m$, (*b*) $-1/m$, (*c*) m, (*d*) $3 + (1/m)$

(*vi*) The mean of a Hypergeometric distribution is

(*a*) $\dfrac{N(M-1)}{N(N-1)}$, (*b*) $\dfrac{M(M-1)}{N(N-1)}$, (*c*) $\dfrac{NM(M-1)}{N(N-1)}$, (*d*) None of these

(*vii*) In a Poisson distribution, the second moment about the origin is 12. Then its third moment about mean is (*a*) 2, (*b*) 3, (*c*) 5, (*d*) 10.

(*viii*) The mean of the binomial distribution ${}^{10}C_x \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{10-x}$; $x = 0, 1, 2, \ldots, 10$ is (*a*) 4, (*b*) 6, (*c*) 5, (*d*) 0.

(*ix*) The mean of Poisson variate is

(*a*) greater than, (*b*) less than, (*c*) equal to, (*d*) twice, its variance.

(*x*) The moment generating function of Geometric distribution is

(*a*) $p(1-qe^t)$, (*b*) $p/(1-qe^t)$, (*c*) $pe^t/(1-qe^t)$, (*d*) None of these.

VIII. By using the uniqueness property of m.g.f.'s, determine the distribution if the M.G.F. is as follows :

(*a*) $M(t) = \left(\frac{1}{2} + \frac{1}{2}e^t\right)^6$, (*b*) $M(t) = \dfrac{(1+e^t)^5}{32}$

(*c*) $M(t) = \dfrac{(1+2e^t)^3}{27}$, (*d*) $M(t) = e^{3(e^t-1)}$,

(*e*) $M(t) = e^{(e^t-1)/4}$, (*f*) $M(t) = \frac{1}{3}e^{-t}\left(e^{-t} - \frac{2}{3}\right)^{-1}$

(*g*) $M(t) = 4(3e^{-t}-1)^{-2}$, (*h*) $M(t) = (3e^{-t}-2)^{-3}$

Ans. (*a*) Binomial, $n = 6, p = \frac{1}{2}$; (*b*) Binomial, $n = 5, p = \frac{1}{2}$;

(*c*) Binomial, $n = 3, p = \frac{2}{3}$; (*d*) Poisson, $\lambda = 3$,

(*e*) Poisson, $\lambda = \frac{1}{4}$, (*f*) Geometric with $p = 1/3$,

(*g*) Negative binomial with $r = 2, p = \frac{2}{3}$;

(*h*) Negative binomial with $r = 3, p = 1/3$.

CHAPTER EIGHT

Theoretical Continuous Distributions

8·1. Rectangular (or Uniform Distribution. A random variable X is said to have a continuous uniform distribution over an interval (a, b) if its probability density function is constant $= k$ (say), over the entire range of X, *i.e.*,

$$f(x) = \begin{cases} k, a < x < b \\ 0, \text{ otnerwise} \end{cases}$$

Since total probability is always unity, we have

$$\int_a^b f(x)\, dx = 1 \;\Rightarrow\; k \int_a^b dx = 1 \quad i.e., \quad k = \frac{1}{b-a}$$

$$\therefore \qquad f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases} \qquad \text{...(8·1)}$$

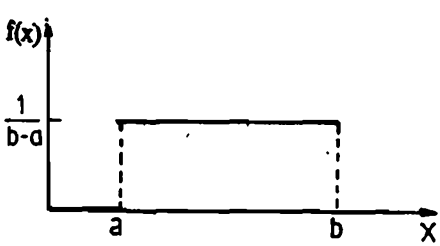
Remarks. 1. a and b, $(a < b)$ are the two parameters of the uniform distribution on (a, b).

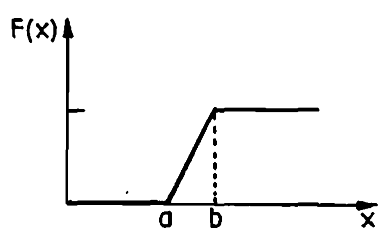
2. The distribution is also known as rectangular distribution, since the curve $y = f(x)$ describes a rectangle over the x-aixs and between the ordinates at $x = a$ and $x = b$.

3. The distribution function $F(x)$ is given by

$$F(x) = \begin{cases} 0, \text{ if } -\infty < x < a \\ \dfrac{x-a}{b-a}, \; a \le x \le b \\ 1, \quad b < x < \infty \end{cases} \qquad \text{...(8·1 } a)$$

Since $F(x)$ is not continuous at $x = a$ and $x = b$, it is not differentiable at these points. Thus $\dfrac{d}{dx} F(x) = f(x) = \dfrac{1}{b-a} \neq 0$, exists everywhere except at the points $x = a$ and $x = b$. and consequently p.d.f. $f(x)$ is given by (8·1).





4. The graphs of uniform p.d.f. $f(x)$ and the corresponding distribution function $F(x)$ are given on page 8·1 :

5. For a rectangular or uniform variate X in $(-a, a)$, p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{1}{2a}, & -a < x < a \\ 0, & \text{otherwise.} \end{cases}$$

8·1·1. Moments of Rectangular Distribution.

$$\mu'_r = \int_a^b x^r f(x)\, dx = \frac{1}{(b-a)} \int_a^b x^r\, dx = \frac{1}{(b-a)} \left[\frac{b^{r+1} - a^{r+1}}{r+1} \right] \quad \text{...(8·2)}$$

In particular

$$\text{Mean} = \mu_1' = \frac{1}{(b-a)} \left[\frac{b^2 - a^2}{2} \right] = \frac{b+a}{2}$$

and

$$\mu_2' = \frac{1}{(b-a)} \left[\frac{b^3 - a^3}{3} \right] = \frac{1}{3} (b^2 + ab + a^2)$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = \frac{1}{3}(b^2 + ab + a^2) - \left(\frac{b+a}{2} \right)^2 = \frac{1}{12} (b-a)^2$$

8·1·2. Moment Generating Function is given by

$$M_X(t) = \int_a^b e^{tx} f(x)\, dx = \frac{e^{bt} - e^{at}}{t(b-a)}$$

8·1·3. Characteristic Function is given by

$$\varphi_X(t) = \int_a^b e^{itx} f(x)\, dx = \frac{e^{ibt} - e^{iat}}{it(b-a)}$$

8·1·4. Mean Deviation about Mean, η is given by

$$\eta = E\,|X - \text{Mean}| = \int_a^b |x - \text{Mean}|\, f(x)\, dx$$

$$= \frac{1}{(b-a)} \int_a^b \left| x - \frac{a+b}{2} \right| dx$$

$$= \frac{1}{(b-a)} \int_{-(b-a)/2}^{(b-a)/2} |t|\, dt \qquad \left[t = x - \frac{a+b}{2} \right]$$

$$= \frac{1}{(b-a)} \cdot 2 \int_0^{(b-a)/2} t\, dt = \frac{b-a}{4}$$

Example 8·1 *If X is uniformly distributed with mean 1 and variance 4/3, find $P(X<0)$.* [Delhi Univ. B.A. (Hons. Spl. Course-Statistics), 1989]

Solution. Let $X \sim U[a, b]$, so that $p(x) = \dfrac{1}{b-a}$; $a < x < b$. We are given:

$$\text{Mean} = \frac{b+a}{2} = 1 \quad \Rightarrow \quad b + a = 2$$

$$\text{Var}(X) = \frac{1}{12}(b-a)^2 = \frac{4}{3} \quad \Rightarrow \quad (b-a)^2 = 16 \quad \Rightarrow \quad b - a = \pm 4$$

Solving, we get : $a = -1$ and $b = 3$; $(a < b)$.

$$\therefore \qquad p(x) = \frac{1}{4} \; ; \; -1 < x < 3$$

$$P(X<0) = \int_{-1}^{0} p(x)\, dx = \frac{1}{4} \left| x \right|_{-1}^{0} = \frac{1}{4}$$

Example 8·2. *Subway trains on a certain line run every half hour between mid-night and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least twenty minutes?*

Solution. Let the *r.v.* X denote the waiting time (in minutes) for the next train. Under the assumption that a man arrives at the station at random, X is distributed uniformly on (0, 30), with *p.d.f.*,

$$f(x) = \begin{cases} \dfrac{1}{30}, & 0 < x < 30 \\ 0, & \text{otherwise} \end{cases}$$

The probability that he has to wait at least 20 minutes is

$$P(X \geq 20) = \int_{20}^{30} f(x)\, dx = \frac{1}{30} \int_{20}^{30} 1 \cdot dx = \frac{1}{30}(30-20) = \frac{1}{3}$$

Example 8·3. *If X has a uniform distribution in $[0, 1]$, find the distribution (p.d.f.) of $-2 \log X$. Identify the distribution also.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1989, '86]

Solution. Let $Y = -2 \log X$. Then the distribution function G of Y is

$$G_Y(y) = P(Y \leq y) = P(-2 \log X \leq y)$$

$$= P(\log X \geq -y/2) = P(X \geq e^{-y/2}) = 1 - P(X \leq e^{-y/2})$$

$$= 1 - \int_0^{e^{-y/2}} f(x)\, dx = 1 - \int_0^{e^{-y/2}} 1 \cdot dx = 1 - e^{-y/2}$$

$$g_Y(y) = \frac{d}{dy} G(y) = \frac{1}{2} e^{-y/2}, \; 0 < y < \infty \qquad ...(*)$$

[$\because$ as X ranges in (0, 1), $Y = -2 \log X$ ranges from 0 to ∞]

Remark. This example illustrates that if $X \sim U[0, 1]$, then $Y = -2 \log X$, has an exponential distribution with parameter $\theta = \frac{1}{2}$. [*c.f.* § 8·6] or $Y = -2 \log X$ has chi-square distribution with $n = 2$ degrees of freedom [*c.f.* Chapter 13, § 13·2].

Example 8·4. *Show that for the rectangular distribution :*

$$f(x) = \frac{1}{2a}, -a < x < a$$

the m.g.f. about origin is $\frac{1}{at}$ *(sinh at). Also show that moments of even order are given by*

$$\mu_{2n} = \frac{a^{2n}}{(2n+1)}$$

Solution. M.G.F. about origin is given by

$$M_X(t) = E(e^{tX}) = \int_{-a}^{a} e^{tx} f(x)\, dx = \frac{1}{2a} \int_{-a}^{a} e^{tx}\, dx$$

$$= \frac{1}{2a} \left| \frac{e^{tx}}{t} \right|_{-a}^{a} = \frac{1}{2at}(e^{at} - e^{-at}) = \frac{\sinh at}{at}$$

$$= \frac{1}{at}\left[at + \frac{(at)^3}{3\,!} + \frac{(at)^5}{5\,!} + \dots \right] = 1 + \frac{a^2 t^2}{3\,!} + \frac{a^4 t^4}{5\,!} + \dots$$

Since there are no terms with odd powers of t in $M(t)$, all moments of odd order about origin vanish, *i.e.*,

$$\mu'_{2n+1} \text{ (about origin)} = 0$$

In particular $\quad \mu_1'$ (about origin) $= 0$, *i.e.*, mean $= 0$

Thus $\quad \mu_r'$ (about origin) $= \mu_r$ $\quad$ (since mean is origin)

Hence $\quad \mu_{2n+1} = 0 \,;\, n = 0, 1, 2, \dots$

i.e., all moments of odd order about mean vanish. The moments of even order are given by

$$\mu_{2n} = \text{coefficient of } \frac{t^{2n}}{(2n)\,!} \text{ in } M(t) = \frac{a^{2n}}{(2n+1)}$$

Example 8·5. *If* X_1 and X_2 *are independent rectangular variates on* $[0, 1]$, *find the distributions of*

(*i*) X_1/X_2, $\quad$ (*ii*) $X_1 X_2$, $\quad$ (*iii*) $X_1 + X_2$, and $\quad$ (*iv*) $X_1 - X_2$

Solution. We are given

$$f_{X_1}(x_1) = f_{X_2}(x_2) = 1 \,;\, 0 < x_1 < 1, 0 < x_2 < 1$$

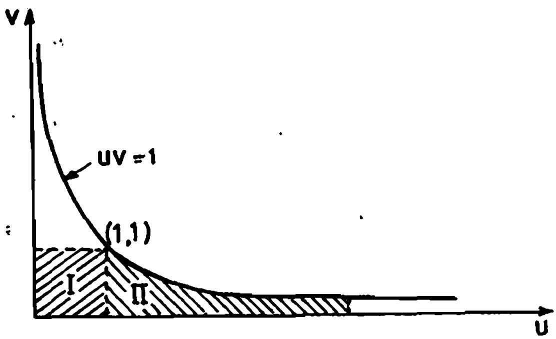
Since X_1 and X_2 are independent, their joint p.d.f. is

$$f(x_1, x_2) = f(x_1) f(x_2) = 1$$

(*i*) Let us transform to

$$u = \frac{x_1}{x_2},\ v = x_2 \quad i.e.,\quad x_1 = uv,\ x_2 = v$$

$$J = \frac{\partial (x_1, x_2)}{\partial (u, v)} = \begin{vmatrix} v & 0 \\ u & 1 \end{vmatrix} = v$$



$x_1 = 0$ maps to $u = 0,\ v = 0$

$x_1 = 1$ maps to $uv = 1$ (Rectangular hyperbola)

$x_2 = 0$ maps to $v = 0$ and $x_2 = 1$ maps to $v = 1$.

The joint p.d.f. of U and V becomes

$$g(u, v) = f(x_1, x_2)\,|J| = v\ ;\ 0 < u < \infty,\ 0 < v < \infty$$

To obtain the marginal distribution of U, we have to integrate out v.

In region (I),

$$g_1(u) = \int_0^1 v\,dv = \left| \frac{v^2}{2} \right|_0^1 = \tfrac{1}{2},\ 0 \le u \le 1$$

In region (II),

$$g_1(u) = \int_0^{1/u} v\,dv = \left| \frac{v^2}{2} \right|_0^{1/u} = \frac{1}{2u^2},\ 1 < u < \infty$$

Hence the distribution of $U = \dfrac{X_1}{X_2}$ is given by

$$g(u) = \tfrac{1}{2},\ 0 \le u \le 1$$

$$= \frac{1}{2u^2},\ 1 \le u < \infty$$

(ii) Let $u = x_1 x_2,\ v = x_1,\ i.e.,\ x_1 = v,\ x_2 = \dfrac{u}{v}$

$$J = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}$$

$x_1 = 0$ maps to $v = 0,\ x_1 = 1$ maps to $v = 1$

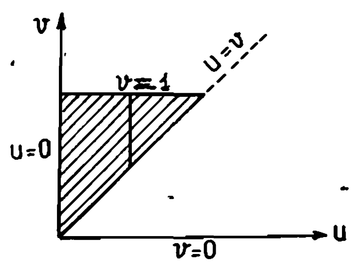
$x_2 = 0$ maps to $u = 0$, and $x_2 = 1$ maps to $u = v$

Moreover, $v = \dfrac{u}{x_2} \Rightarrow v \geq u$ (since $0 < x_2 < 1$),

The joint p.d.f. of U and V is

$$g(u, v) = f(x_1, x_2) \mid J \mid = \frac{1}{v} \,;\, 0 < u < 1, 0 < v < 1$$

$$g(u) = \int_u^1 \frac{1}{v}\, dv = [\log v]_u^1 = -\log u, \; 0 < u < 1$$



(*iii*) and (*iv*). Let $u = x_1 + x_2$,
$v = x_1 - x_2$

i.e., $x_1 = \dfrac{u+v}{2}$, $x_2 = \dfrac{u-v}{2}$

$x_1 = 0 \Rightarrow u + v = 0$ i.e., $v = -u$
$x_2 = 0 \Rightarrow u - v = 0$ i.e., $v = u$
$x_1 = 1 \Rightarrow u + v = 2$
$x_2 = 1 \Rightarrow u - v = 2$

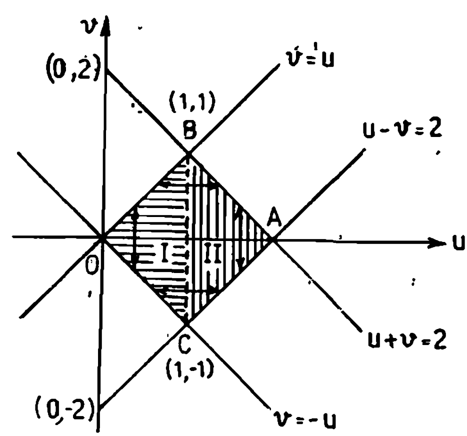
and $J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$

$\therefore \; g(u, v) = f(x_1, x_2) \mid J \mid = \frac{1}{2}, \; 0 < u < 2, -1 < v < 1$

In region (I), (see figure below)

$$g_1(u) = \int_{-u}^{u} \frac{1}{2}\, dv = \frac{1}{2} \Big| v \Big|_{-u}^{u} = u$$

and in region (II),



$$g_2(u) = \int_{u-2}^{2-u} \tfrac{1}{2}\, dv = \tfrac{1}{2} \mid v \mid_{u-2}^{2-u} = 2 - u$$

$$\therefore \quad g(u) = \begin{cases} u, & 0 < u < 1 \\ 2 - u, & 1 < u < 2 \end{cases}$$

For the distribution of V, we split the region as: OAB and OAC

In region OAB :

$$h_1(v) = \int_{v}^{2-v} \tfrac{1}{2}\, du = \tfrac{1}{2}[\,2 - v - v\,] = 1 - v, \; 0 < v < 1$$

In region OAC :

$$h_2(v) = \int_{-v}^{2+v} \tfrac{1}{2}\, du = \tfrac{1}{2}[\,2(1+v)\,] = 1 + v, \; -1 < v < 0$$

Hence the distribution of $V = X_1 - X_2$ is given by

$$h(v) = \begin{cases} 1 - v, & 0 < v < 1 \\ 1 + v, & -1 < v < 0 \end{cases}$$

Example 8·6. *If X is a random variable with a continuous distribution function F, then $F(X)$ has a uniform distribution on [0, 1].*

[Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1987,' 85]

Solution. Since F is a distribution function, it is non-decreasing. Let $Y = F(X)$, then the distribution function G of Y is given by

$$G_Y(y) = P(Y \le y) = P[\,F(X) \le y\,] = P[\,X \le F^{-1}(y)\,],$$

the inverse exists, since F is non-decreasing and given to be continuous.

$$\therefore \quad G_Y(y) = F[\,F^{-1}(y)\,],$$

since F is the distribution function of X.

$$\therefore \quad G_Y(y) = y$$

Therefore the p.d.f. of $Y = F(X)$ is given by:

$$g_Y(y) = \frac{d}{dy}[\,G_Y(y)\,] = 1$$

Since F is a d.f., Y takes the values in the range [0, 1].

Hence $g_Y(y) = 1, 0 \le y \le 1$

$\Rightarrow$ Y is a uniform variate on $[\,0, 1\,]$.

Remark. Suppose X is a random variable with p.d.f.,

$$f_X(x) = \begin{cases} e^{-x}, & x \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

then

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - e^{-x}, & \text{if } x \ge 0 \end{cases}$$

Then by above result $F(X) = 1 - e^{-X}$ is uniformly distributed on [0, 1].

Example 8·7. *If X and Y are independent rectangular variates for the range $-a$ to a each, then show that the sum $X + Y = U$, has the probability density*

$$\varphi(u) = \frac{2a+u}{4a^2}, \; -2a \le u \le 0$$

$$\varphi(u) = \frac{2a-u}{4a^2}, \; 0 \le u \le 2a$$

Solution. Since X and Y are independent rectangular variates, each in the interval $(-a, a)$, we have

$$f_1(x) = \begin{cases} \frac{1}{2a}, & -a < x < a \\ 0, & \text{elsewhere} \end{cases}$$

and
$$f_2(y) = \begin{cases} \frac{1}{2a}, & -a < y < a \\ 0, & \text{elsewhere} \end{cases}$$

Hence by compound probability theorem, the joint probability differential of X and Y is given by

$$dP(x, y) = f_1(x) f_2(y)\, dx\, dy = \frac{1}{4a^2}\, dx\, dy, \; -a < (x, y) < a$$

Let us define new variables U and V as follows :

$$u = x + y, \qquad v = x - y$$

$$\Rightarrow \qquad x = \frac{u+v}{2} \quad \text{and} \quad y = \frac{u-v}{2}$$

Jacobian of the transformation J is given by

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

Thus the probability differential of U and V becomes

$$dG(u, v) = \frac{1}{4a^2} |J|\, du\, dv = \frac{1}{8a^2}\, du\, dv \qquad \text{...(*)}$$

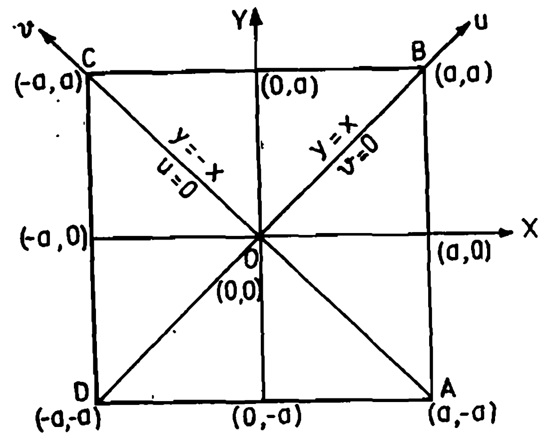
Integrating w.r.to. v over specified range, we can find the distribution of U.

Let us consider the region to the left of $v-$ axis, *i.e.*, to the left of the line AC. In this region, the values of v are bounded by the lines $x = -a$ and $y = -a$.

For fixed values of u,

$$x = -a \quad \Rightarrow \quad \frac{u+v}{2} = -a \quad \Rightarrow \quad v = -(u + 2a)$$

and
$$y = -a \quad \Rightarrow \quad \frac{u-v}{2} = -a \quad \Rightarrow \quad v = (u + 2a)$$



Thus integrating (*) w.r.to. v between the limits $-(u+2a)$ and $(u+2a)$, the distribution of U becomes

$$g_1(u)\,du = \int_{-(u+2a)}^{u+2a} \frac{1}{8a^2}\cdot du\,dv = \frac{1}{8a^2}\,|\,v\,|_{-(u+2a)}^{u+2a}\,du = \frac{u+2a}{4a^2}\,du$$

In the region to the left of v-axis, *i.e.*, below the line AC, u varies from the points $(x=-a,\ y=-a)$ to the point $(x=0, y=0)$ and since $u=x+y$, in this region u lies between $(-a-a)$ and $(0+0)$, *i.e.*, between $-2a$ to 0.

$$\therefore \qquad g_1(u)\,du = \frac{u+2a}{4a^2}, \quad -2a \le u \le 0$$

In the region to the right of v–axis, *i.e.*, above the line AC, the values of v are bounded by the lines $x=a$ and $y=a$ and for fixed values of u,

$$x=a \quad \Rightarrow \quad \frac{u+v}{2}=a \quad \Rightarrow \quad v = 2a-u$$

$$y=a \quad \Rightarrow \quad \frac{u-v}{2}=a \quad \Rightarrow \quad v=-(2a-u)$$

In this region u varies from the point $(x=0, y=0)$ to the point $(x=a, y=a)$, i.e., $u=x+y$ varies from 0 to $2a$. Thus integrating (*) w.r.to. v between the limits $-(2a-u)$ to $(2a-u)$, we get the distribution of U as

$$g_1(u)\,du = \int_{-(2a-u)}^{2a-u} \frac{1}{8a^2}\,du\,dv = \frac{1}{8a^2}\,|\,v\,|_{-(2a-u)}^{2a-u}\,du$$

$$= \frac{2a-u}{4a^2}\,du, \ 0 \le u \le 2a$$

For an alternative and simpler solution, see Remark 5 to § 8·1·5, (Triangular Distribution).

Example. 8·8. *On the x-axis $(n+1)$ points are taken independently between the origin and $x=1$, all positions being equally likely. Show that probability* *that the* $(k+1)$ *th of these points, counted from the origin, lies in the interval* $x - \frac{1}{2}\,dx$ *to* $x + \frac{1}{2}\,dx$ *is*

$$\binom{n}{k}(n+1)\,x^k\,(1-x)^{n-k}\,dx$$

Verify that integral of this expression from $x = 0$ *to* $x = 1$ *is unity.*

Solution. Here X is given to be a random variable uniformly distributed on $[\,0, 1\,]$.

$$\therefore \qquad f_X(x) = 1,\ 0 \le x \le 1$$

Now $$P(0 < X < x) = \int_0^x f(x)\,dx = \int_0^x 1\,.\,dx = x \qquad \text{...(1)}$$

$$\therefore \qquad P(X > x) = 1 - P(X \le x) = 1 - x \qquad \text{...(2)}$$

Also $$P\left(x - \frac{dx}{2} < X < x + \frac{dx}{2}\right) = \int_{x-\frac{dx}{2}}^{x+\frac{dx}{2}} f(x)\,dx = dx \qquad \text{...(3)}$$

Required probability 'p' is given by

$p = P\,\Big\{$out of $(n+1)$ points, k points lie in the closed interval $\left[\,0, x - \frac{dx}{2}\right]$

and out of the remaining $(n+1-k)$ points, $(n-k)$ points lie in

$\left[\,x + \frac{dx}{2}, 1\,\right]$ and one point lies in $\left(x - \frac{dx}{2}, x + \frac{dx}{2}\right)\Big\}$

$$= \left[\binom{n+1}{k}x^k\right] \times \left[\binom{n+1-k}{n-k}(1-x)^{n-k}\right] \times dx,$$

on using (1), (2) and (3) respectively.

$$\therefore \qquad p = \frac{(n+1)\,!}{k\,!\,(n+1-k)\,!}\,.\,x^k\,.\,\frac{(n+1-k)\,!}{(n-k)\,!}\,.\,(1-x)^{n-k}\,dx$$

$$= \binom{n}{k}(n+1)\,x^k\,(1-x)^{n-k}\,dx$$

To prove that the area of this expression from $x = 0$ to $x = 1$ is unity, use Beta-integral

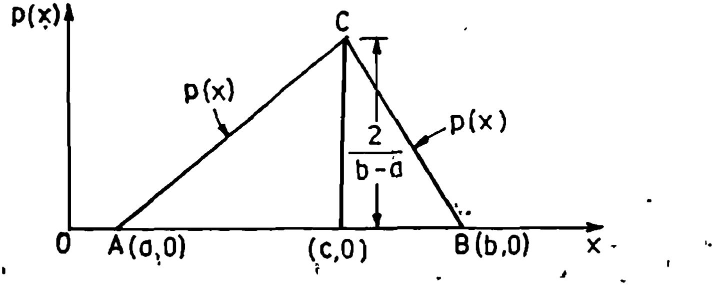
$$\int_0^1 x^{m-1}\,(1-x)^{n-1}\,dx = B\,(m, n) = \frac{\Gamma m\ \Gamma n}{\Gamma\,(m+n)}\,;\ m > 0, n > 0\,.$$

8·1·5. Triangular Distribution. A random variable X is said to have a triangular distribution in the interval (a, b), if its p.d.f. is given by:

$$f(x) = \begin{cases} 2\,(x-a)/\left\{(b-a)\,(c-a)\right\} & ;\ a < x \le c \\ 2\,(b-x)/\left\{(b-a)\,(b-c)\right\} & ;\ c < x < b \end{cases} \qquad \text{...(8·2a)}$$

Remarks. 1. We write $X \sim$ Trg. (a, b), with peak at $x = c$. The graph of the **p.d.f. is shown in the diagram on page 8·11.**

2. The distribution is so called because the graph of its p.d.f. is a triangle with peak at $x = c$.



3. The m.g.f. of Trg (a, b) variate, with peak at $x = c$ is given by:

$$M_X(t) = \int_a^b e^{tx} f(x)\, dx = \left(\int_a^c + \int_c^b \right) e^{tx} f(x)\, dx.$$

$$= \frac{2}{(b-a)(c-a)} \int_a^c e^{tx}(x-a)\, dx + \frac{2}{(b-a)(b-c)} \int_c^b e^{tx} . (b-x)\, dx$$

$$= \frac{2}{t^2} \left\{ \frac{e^{at}}{(a-b)(a-c)} + \frac{e^{ct}}{(c-a)(c-b)} + \frac{e^{bt}}{(b-a)(b-c)} \right\} ; a < b < c$$

(On integration by parts) ...(8·2b)

4. In particular, taking $a = 0$, $c = 1$ and $b = 2$, in (8·2a), the p.d.f. of the Trg (0, 2) variate with peak at $x = 1$ is given by:

$$f(x) = \begin{cases} x\,; & 0 \le x \le 1 \\ 2 - x\,; & 1 \le x \le 2 \\ 0, & \text{otherwise} \end{cases} \qquad \text{...(8·2 } c)$$

and its m.g.f. is, $M_X(t) = (e^t - 1)^2 / t^2$, ...(8·2$d$)

which is left as an exercise to the reader.

5. In particular, replacing a by $-2a$, b by $2a$ and c by 0, the p.d.f. of triangular distribution on the interval $(-2a, 2a)$ with peak at $x = 0$ is given by:

$$f(x) = \begin{cases} (2a + x)/4a^2\,; & -2a < x < 0 \\ \\ (2a - x)/4a^2\,; & 0 < x < 2a \end{cases} \qquad \text{...(8·2}e)$$

The m.g.f. of (8·2e) is given by :

$$M_X(t) = \int_{-2a}^{2a} e^{tx} f(x)\, dx$$

$$= \frac{1}{4a^2} \left[\int_{-2a}^{0} e^{tx} . (2a + x)\, dx + \int_0^{2a} e^{tx} (2a - x)\, dx \right]$$

$$= \frac{1}{4a^2}\left[e^{tx}\left\{ \frac{2a+x}{t} - \frac{1}{t^2} \right\} \right]_{-2a}^{0} + \frac{1}{4a^2}\left[e^{tx}\left\{ \frac{2a-x}{t} + \frac{1}{t^2} \right\} \right]_{0}^{2a}$$

[On integrating by parts]

$$= \frac{1}{4a^2}\left[-\frac{2}{t^2} + \frac{1}{t^2}\left\{ e^{2at} + e^{-2at} \right\} \right]$$

$$= \frac{1}{4a^2 t^2}\left\{ e^{2at} + e^{-2at} - 2 \right\} = \left[\frac{e^{at} - e^{-at}}{2\,at} \right]^2 \qquad ...(8{\cdot}2\,f)$$

Aliter. We may obtain (8·2 f) directly from (8·2 b) on replacing a by $-2a$, b by $2a$ and c by 0.

Example 8·9. *If X and Y are i.i.d. $U[-a, a]$ variates, find the p.d.f. of $Z = X + Y$ and identify the distribution.*

Solution. Since X and Y are i.i.d. $U[-a, a]$, we have : [*c.f.* § 8·1·2.],

$$M_X(t) = M_Y(t) = (e^{at} - e^{-at})/(2\,at) \qquad ...(*)$$

$$M_{X+Y}(t) = M_X(t)\, M_Y(t) = \left[\frac{e^{at} - e^{-at}}{2at} \right]^2, \qquad ...(**)$$

since X and Y are independent.

But (**) is the m.g.f. of Trg $(-2a, 2a)$ variate with peak at $x = 0$

[*c.f.* Remark 5, equation (8·2f)]

Hence by uniqueness theorem of m.g.f., $Z = X + Y \sim$ Trg $(-2a, 2a)$ with p.d.f. as given in (8·2 e), Remark 5.

Aliter $M_{X+Y}(t) = \dfrac{1}{4a^2 t^2}\left[e^{2at} - 2 + e^{-2at} \right]$ [From (**)]

$$= \frac{2}{t^2}\left[\frac{e^{-2at}}{(-2a-0)(-2a-2a)} + \frac{e^{0.t}}{(0+2a)(0-2a)} + \frac{e^{2at}}{(2a-0)(2a+2a)} \right]$$

which is of the form (8·2 b), [*c.f.* Remark 3], with a replaced by $-2a$ and b replaced by $2a$ and c by 0. Hence $X + Y \sim$Trg $(-2a, 2a)$ with p.d.f. $p(x)$ given in (8·2e).

Remarks 1. The distribution of $X + Y$ has also been obtained in Example 8·7.

2. Similarly we can find the distribution of $X - Y$.

$$M_{X-Y}(t) = M_X(t)\,.\,M_Y(-t) = \left[\frac{e^{at} - e^{-at}}{2at} \right]^2 \qquad \text{[From (*)]}$$

$\Rightarrow X - Y \sim$Trg $(-2a, 2a)$, with peak at $x = 0$.

EXERCISE 8 (a)

1. The bus company A schedules a north bound bus every 30 minutes at a certain bus-stop. A man comes to the stop at a random time. Let the random variable X count the number of minutes he has to wait for the next bus. Assume X has a uniform distribution over the interval (0, 30). This is how we interpret the statement that he enters the station at the random time].

(*i*) For each $k = 5, 10, 15, 20, 30$ compute the probability that he has to wait at least k minutes for the next bus.

(*ii*) A competitor, the bus company B is allowed to schedule a north bound bus every 30 minutes at the same station but at least 5 minutes must elapse between the arrivals of the competitive buses. Assume the passengers come at the bus stop at random times and always board the first bus that arrives. Show that the company B can arrange its schedule so that it receives five times as many passengers as that of its competitor.

2. (*a*) A random variable X has a uniform distribution over $(-3, 3)$, compute

(*i*) $P(X = 2)$, $P(X < 2)$, $P(|X| < 2)$ and $P(|X - 2| < 2)$

(*ii*) Find k for which $P(X > k) = 1/3$. **[Gorakhpur Univ. B.Sc. 1992)**

(*b*) Suppose that X is uniformly distributed over $(-\alpha, +\alpha)$, where $\alpha > 0$. Determine α so that

(*i*) $P(X > 1) = 1/3$, (*ii*) $P(X < 1/2) = 0{\cdot}3$ and

(*ii*) $P(|X| < 1) = P(|X| > 1)$.

Ans. (*i*) $\alpha = 3$, (*ii*) $\alpha = 5/6$, (*iii*) $\alpha = 2$.

(*c*) Calculate the coefficient of variation for the rectangular distribution in $(0, b)$ given that the probability law of the distribution is

$$P(X \le t) = \frac{t}{b}$$

(*d*) If X is uniformly distributed over $[1, 2]$, find z so that

$$P(X > z + \mu_x) = \frac{1}{4} \qquad \textbf{(Ans. } Z = \frac{1}{4}\textbf{)}.$$

3 (*a*). If a random variable X has the density function $f(x)$, prove that

$$Y = \int_{-\infty}^{x} f(x)\, dx$$

has a rectangular distribution over (0, 1). If

$$f(x) = \tfrac{1}{2}(x - 1), \quad 1 \le x \le 3$$
$$= 0, \text{ otherwise}$$

determine what interval for Y will correspond to the interval

$$1{\cdot}1 \le X \le 2{\cdot}9.$$

Ans. $y = F(x) = (x-1)^2/4$; $1 \le x \le 3$; $0{\cdot}0025 \le y \le 0{\cdot}9025$

(*b*). Show that whatever be the distribution function $F(x)$ of a r.v. X,

$$P[a \le F(X) \le b] = b - a, \quad 0 \le (a, b) \le 1.$$

[Delhi Univ. B.Sc. (Stat. Hons), 1986]

Hint. $Y = F(X) \sim U[0, 1]$.

4. (*a*) For the rectantular distribution,

$$f(x) = \frac{1}{2a}\ ;\ -a \le x \le a\ .$$
$$= 0,\ \text{otherwise},$$

show that the moments of odd order are zero, and $\mu_{2r} = a^{2r}/(2r+1)$.

[Madurai Kamraj Univ. B.Sc. 1992]

(*b*) A distribution is given by

$$f(x)\,dx = \frac{1}{2a}\,dx,\ -a \le x \le a$$

Find the first four central moments and obtain β_1 and β_2.

[Delhi Univ B.Sc. Oct., 1992; Madras Univ. B.Sc., 1991]

(*c*) For a rectangular distribution

$$dP = k\,dx,\ 1 \le x \le 2,$$

show that Arithmetic mean > Geometric mean > Harmonic mean.

[Vikram Univ. B.Sc. 1993)

(*d*) If the random variable X follows the rectangular distribution with p.d.f.,

$$f(x) = 1/\theta,\ 0 \le x \le \theta\ ,$$

derive the first four moments and the skewness and kurtosis conffîcients of the distribution.

(*e*) Let X and Y be independent variates which are uniformly distributed over the unit interval (0,1). Find the distribution function and the p.d.f. of random variable $Z = X + Y$. Is Z a uniformly distributed variable ? Give reasons.

[Delhi Univ. B.Sc. (Maths. Hons.), 1986]

5. Let X_1 and X_2 be independent random variables unifromly distributed over the interval (0, 1). Find

(*i*) $P(X_1 + X_2 < 0{\cdot}5)$, (*ii*) $P(X_1 - X_2 < 0{\cdot}5)$,

(*iii*) $P(X_1^2 + X_2^2 < 0{\cdot}5)$, (*iv*) $P(e^{-X_1} < 0{\cdot}5)$, and (*v*) $P(\cos \pi X_2 < 0{\cdot}5)$.

Ans. (*i*) 0·125, (*ii*) 0·875. (*iii*) 0·393, (*iv*) $1 - \log 2$, and (*v*) $2/3$.

6. A random variable X is uniformly distributed over (0, 1), find the probability density functions of

(*i*) $Y = X^2 + 1$, and (*ii*) $Z = 1/(X+1)$.

7. (*a*) If the random variable X is uniformly distributed over $(0, \frac{1}{2}\pi)$, compute the expectation of the function $\sin X$. Also find the distribution of $Y = \sin X$, and show that the mean of this distribution is the same as the above expectation.

Ans. $2/\pi$, $f_Y(y) = 2/(\pi\sqrt{1-y^2}),\ 0 < y < 1$.

(*b*) If $X \sim U[-\pi/2, \pi/2]$ distributed, find the p.d.f. of $Y = \tan X$.

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1989]

8. (*a*) Show that for the rectangular distribution :

$$dF = dx,\ 0 \le x < 1$$

μ'_1 (about origin) = 1/2, variance = 1/12 and mean deviation about mean = 1/4. **[Madras Univ B.Sc. Sept. 1991; Delhi U. B.Sc. Sept. 1992]**

(*b*) Find the characteristic function of the random variable $Y = \log F(X)$ where $F(X)$ is the distribution function of a random variable X. Evaluate the rth moment of Y.

9. If $X \sim U[0, 1]$, find the distribution of $Y = 1/X$. Find $E(1/X)$, if it exists.

Ans. $g_Y(y) = 1/y^2$; $1 \le y < \infty$; $E(Y) = E(1/X)$ does not exist.

10. Let X be uniformly distributed on $[-1, 1]$. Find the distribution function and hence the p.d.f. of $Y = X^2$. **[Delhi Univ. B.Sc. (Maths. Hons.), 1988]**

11. Let $f_X(x) = 6x(1-x)$; $0 \le x \le 1$. Find y as a function of x such that Y has p.d.f.

$$g(y) = 3(1 - \sqrt{y})\,; \quad 0 \le y \le 1$$

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1988]

Hint.
$$F(x) = \int_0^x f(x)\,dx = 3x^2 - 2x^3 \sim U[0, 1]$$

$$G(y) = \int_0^y g(y)\,dy = 3y - 2y^{3/2} \sim U[0, 1]$$

Setting $F(x) = G(y)$, we get $y = x^2$.

12. The variates a and b are independently and uniformly distributed in the intervals [0,6] and [0, 9] respectively. Find the probability that $x^2 - ax + b = 0$ has two real roots.

Ans. $P(b \le a^2/4) = \int_{a=0}^{1} \int_{b=0}^{a^2/4} \dfrac{1}{6 \times 9}\, da\, db = 1/3$.

13. Find the probability that the roots of the equation $x^2 + 2bx + c = 0$ should be real, given that $b \sim U[-\alpha, \alpha]$ and $c \sim U[-\beta, \beta]$ are independent.

Ans. Probability $= P(b^2 \ge c) = 1 - P(b^2 \le c) = 1 - P(|b| \le \sqrt{c})$

$$= \int_{-\beta}^{\beta} \left(\int_{-\sqrt{c}}^{\sqrt{c}} \left(\frac{1}{2\alpha}\right)\left(\frac{1}{2\beta}\right) db \right) dc$$

14. If a, b, c are randomly chosen between 0 and 1, find the probability that the quadratic equation $ax^2 + bx + c = 0$ has real roots.

Ans. Probability $= P(b^2 \ge 4ac) = 1 - \int_0^1 \int_0^1 \int_{b=0}^{\sqrt{4ac}} 1 \, da\, dc\, db = \dfrac{8}{9}$

15. (*a*) Suppose X has a rectangular distribution on $(-1, 1)$. Compute $P\left[\dfrac{|X - E(X)|}{\sigma_X} \ge 2\right]$ and compare it with the upper bound given by Chebyshev's inequality.

(*b*) Compare the upper bound of the probability,

$$P\left\{|X-E(X)|\geq 2\sqrt{V(X)}\right\},$$

obtained from Chebyshev's inequality with exact probability if X is uniformly distributed over $(-1, 3)$.

Ans. (*b*) Probability $\leq 1/4$, Exact Probability $= 0$

16. Two independent variates are each uniformly distributed within the range $-a$ to $+a$. Show that their sum X has a probability density given by

$$f(x) = \frac{2a+x}{4a^2}, \quad -2a \leq x \leq 0$$

$$= \frac{2a-x}{4a^2}, \quad 0 \leq x \leq 2a$$

Verify that the m.g.f. calculated from the value of $f(x)$ is equal to

$$\left(\frac{1}{at}\sinh at\right)^2$$

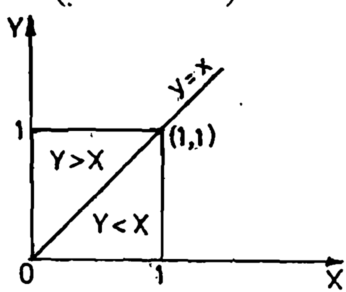
17. The random variables X and Y are independent and both have the uniform distribution on $[0, 1]$. Let $Z = |X-Y|$. Prove that, for real θ,

$$\varphi(Z, \theta) = 2\left[1+i\theta-e^{i\theta}\right]/2.$$

Hence deduce the general expression for $E(Z^n)$.

Hint. $\varphi(\theta\,;|X-Y|) = \int_0^1\int_0^1 e^{i\theta|x-y|} f(x,y)\,dx\,dy$

$$= 2\int_0^1\left(\int_0^x e^{i\theta(x-y)}\,dy\right)dx$$



Ans. $2/[(n+1)(n+2)]$

18. If X and Y are independently and uniformly distributed random variables in the interval $(0, 1)$, show that the distribution of $X+Y$ is given by the density function

$$f(z) = \begin{cases} z & 0 \leq z < 1 \\ 2-z & 1 \leq z \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

[Hint. See Triangular distribution]

19. Ship A makes radio signals to the base and the probability of the interval between consecutive signals is uniformly distributed between 4 hours and 24 hours and is zero outside this range. Ship B makes radio signals to the base and the probability of the interval between consecutive signals is uniformly distributed between 10 hours and 15 hours and is zero outside this range.

(*i*) Ship A has just signalled. What is the probability that it will make two further signals in the next 12 hours ?

(*ii*) Ships A and B have just signalled at the same time. What is the probability that Ship A will make at least two further signals before ship B next signals? **[Institute of Actuaries (London), April 1978]**

20. If $X \sim U[0, 1]$, prove that for $b < c$ fixed, $Y = (c-b)X + b$ is uniform on $[b, c]$.

8·2. Normal Distribution. The normal distribution was first discovered in 1733 by English mathematician De-Moivre, who obtained this continuous distribution as a limiting case of the binomial distribution and applied it to problems arising in the game of chance. It was also known to Laplace, no later than 1774 but through a historical error it was credited to Gauss, who first made reference to it in the beginning of 19th century (1809), as the distribution of errors in Astronomy. Gauss used the normal curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies. Throughout the eighteenth and nineteeth centuries, various efforts were made to establish the normal model as the underlying law ruling all continuous random variables. Thus, the name "*normal*". These efforts, however, failed because of false premises. The normal model has, nevertheless, become the most important probability model in statistical analysis.

Definition. *A random variable X is said to ahve a normal distribution with parameters μ (called "mean") and σ^2 (called "variance") if its density function is given by the probability law :*

$$f(x\,;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2\right]$$

or $$f(x\,;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2},$$

$$-\infty < x < \infty,\ -\infty < \mu < \infty,\ \sigma > 0 \quad ...(8\cdot3)$$

Remarks. 1. A random variable X with mean μ and variance σ^2 and following the normal law (8·3) is expressed by $X \sim N(\mu, \sigma^2)$

2. If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma}$, is a standard normal variate with

$$E(Z) = 0 \text{ and } \operatorname{Var}(Z) = 1$$

and we write $Z \sim N(0,1)$.

3. The p.d.f. of standard normal variate Z is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}, \; -\infty < z < \infty$$

and the corresponding distribution function, denoted by $\Phi(z)$ is given by

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \varphi(u)\, du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du$$

We shall prove below two important results on the distribution function $\Phi(.)$ of standard normal variate.

Result 1. $\Phi(-z) = 1 - \Phi(z)$

Proof. $\Phi(-z) = P(Z \le -z) = P(Z \ge z)$ (By symmetry)

$$= 1 - P(Z \le z)$$

$$= 1 - \Phi(z)$$

Result 2. $P(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$, where $X \sim N(\mu, \sigma^2)$

Proof. $P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$; $\left(Z = \frac{X-\mu}{\sigma}\right)$

$$= P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

4. The graph of $f(x)$ is a famous *'bell-shaped'* curve. The top of the bell is directly above the mean μ. For large values of σ, the curve tends to flatten out and for small values of σ, it has a sharp peak.

8·2·1. Normal Distribution as a Limiting form of Binomial Distribution. Normal distribution is another limiting form of the binomial distribution under the following conditions :

(*i*) n, the number of trials is indefinitely large, *i.e.*, $n \to \infty$ and

(*ii*) neither p nor q is very small.

The probability function of the binomial distribution with parameters n and p is given by

$$p(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!\,(n-x)!}\, p^x q^{n-x} \; ; \; x = 0, 1, 2, \ldots, n \quad \ldots(*)$$

Let us now consider the standard binomial variate :

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - np}{\sqrt{npq}} \; ; \; X = 0, 1, 2, \ldots, n \quad \ldots(**)$$

When $X = 0,\; Z = \frac{-np}{\sqrt{npq}} = -\sqrt{np/q}$

and when $X = n$, $Z = \dfrac{n - np}{\sqrt{npq}} = \sqrt{nq/p}$

Thus in the limit as $n \to \infty$, Z takes the values from $-\infty$ to ∞. Hence the distribution of X will be a continuous distribution over the range $-\infty$ to ∞.

We want the limiting form of (*) under the above two conditions. Using Stirling's approximation to $r\,!$ for large r, *viz.*,

$$\lim_{r \to \infty} r\,! \simeq \sqrt{2\pi}\, e^{-r}\, r^{r + (1/2)},$$

we have in the limit as $n \to \infty$ and consequently $x \to \infty$,

$$\lim p(x) = \lim \left[\frac{\sqrt{2\pi}\, e^{-n}\, n^{n+\frac{1}{2}}\, p^x\, q^{n-x}}{\sqrt{2\pi}\, e^{-x}\, x^{x+\frac{1}{2}}\, \sqrt{2\pi}\, e^{-(n-x)}\, (n-x)^{n-x+\frac{1}{2}}} \right]$$

$$= \lim \left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \cdot \frac{(np)^{x+\frac{1}{2}}\,(nq)^{n-x+\frac{1}{2}}}{x^{x+\frac{1}{2}}\,(n-x)^{n-x+\frac{1}{2}}} \right]$$

$$= \lim \left[\frac{1}{\sqrt{2\pi}\,\sqrt{npq}} \left(\frac{np}{x} \right)^{x+\frac{1}{2}} \left(\frac{nq}{n-x} \right)^{n-x+\frac{1}{2}} \right] \qquad \ldots(***)$$

From (**), we have

$$X = np + Z\sqrt{npq} \;\Rightarrow\; \frac{X}{np} = 1 + Z\,\sqrt{q/(np)}$$

Also

$$n - X = n - np - Z\sqrt{npq} = nq - Z\sqrt{npq}$$

$$\therefore \quad \frac{n - X}{nq} = 1 - Z\,\sqrt{p/(nq)}\,. \quad \text{Also } dz = \frac{1}{\sqrt{npq}}\,dx$$

Hence the probability differential of the distribution of Z, in the limit is given from (***) by

$$d\,G(z) = g(z)\,dz = \lim_{n \to \infty} \left[\frac{1}{\sqrt{2\pi}} \times \frac{1}{N} \right] dz \qquad \ldots(8\cdot4)$$

where $\quad N = \left[\dfrac{x}{np} \right]^{x+\frac{1}{2}} \left[\dfrac{n-x}{nq} \right]^{n-x+\frac{1}{2}}$

$$\log N = (x + \tfrac{1}{2}) \log (x/np) + (n - x + \tfrac{1}{2}) \log \{ (n - x)/nq \}$$

$$= (np + z\sqrt{npq} + \tfrac{1}{2}) \log [\, 1 + z\sqrt{(q/np)}\,]$$

$$+ (nq - z\sqrt{npq} + \tfrac{1}{2}) \log [\, 1 - z\sqrt{(p/nq)}\,]$$

$$= (np + z\sqrt{npq} + \tfrac{1}{2}) \left[z \,.\, \sqrt{(q/np)} - \tfrac{1}{2} z^2 (q/np) + \tfrac{1}{3} z^3 (q/np)^{3/2} - \ldots \right]$$

$$+ (nq - z\sqrt{npq} + \tfrac{1}{2}) \left[- z\sqrt{(p/nq)} - \tfrac{1}{2} z^2 (p/nq) - \tfrac{1}{3} z^3 (p/nq)^{3/2} - \ldots \right]$$

$$= \Bigg[\Bigg\{ z\sqrt{npq} - \tfrac{1}{2} qz^2 + \tfrac{1}{3} z^3 \frac{q^{3/2}}{\sqrt{np}} + z^2 q - \tfrac{1}{2} z^3 \frac{q^{3/2}}{\sqrt{np}} + \tfrac{1}{2} z\sqrt{q/np} - \tfrac{1}{4} z^2 \frac{q}{np} + \ldots \Bigg\}$$

$$+ \Bigg\{ - z\sqrt{npq} - \tfrac{1}{2} z^2 p - \tfrac{1}{3} z^3 \frac{p^{3/2}}{\sqrt{nq}} + z^2 p + \tfrac{1}{2} z^3 \frac{p^{3/2}}{\sqrt{np}} - \tfrac{1}{2} z\sqrt{p/nq} - \tfrac{1}{4} z^2 \frac{p}{np} + \ldots \Bigg\} \Bigg]$$

i.e.,

$$\log N = \left[-\frac{1}{2} z^2 (p+q) + z^2 (p+q) + \frac{z}{2\sqrt{n}} \left\{ \sqrt{q/p} + \sqrt{p/q} \right\} + 0 \left\{ n^{-1/2} \right\} \right]$$

$$= \frac{z^2}{2} + 0\,(n^{-1/2}) \to \frac{z^2}{2} \text{ as } n \to \infty$$

$$\therefore \quad \lim_{n \to \infty} \log N = \frac{z^2}{2} \quad \Rightarrow \quad \lim_{n \to \infty} N = e^{z^2/2}$$

Substituting in (8·4), we get

$$d\,G(z) = g\,(z)\,dz = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \; -\infty < z < \infty \qquad \ldots(8·4\,a)$$

Hence the probability function of Z is

$$g\,(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \; -\infty < z < \infty \qquad \ldots(8·4\,b)$$

This is the probability density function of the *normal distribution* with mean 0 and unit variance.

If X is normal variate with mean μ and s.d. σ then $Z = (X - \mu)/\sigma$ is standard normal variate. Jacobian of transformation is $1/\sigma$. Hence substituting in {8·4 (*b*)}, the p.d.f. of a normal variate X with $E\,(X) = \mu$, Var $(X) = \sigma^2$ is given by

$$f_X(x) = \begin{cases} \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, & -\infty < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

Remark. Normal distribution can also be obtained as a limiting case of Poisson Distribution with the parameter $\lambda \to \infty$.

8·2·2. Chief Characteristics of the Normal Distribution and Normal Probability Curve. The normal probability curve with mean μ and standard deviation σ is given by the equation

$$f\,(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \; -\infty < x < \infty$$

and has the following properties :

(*i*) The curve is bell shaped and symmetrical about the line $x = \mu$.

(*ii*) Mean, median and mode of the distribution coincide.

(*iii*) As x increases numerically, $f(x)$ decreases rapidly, the maximum probability occurring at the point $x = \mu$, and given by $[p(x)]_{max} = \dfrac{1}{\sigma\sqrt{2\pi}}$.

(*iv*) $\beta_1 = 0$ and $\beta_2 = 3$.

(*v*) $\mu_{2r+1} = 0, (r = 0, 1, 2, \ldots)$,
and $\mu_{2r} = 1.3.5 \ldots (2r-1)\sigma^{2r}, (r = 0, 1, 2, \ldots)$.

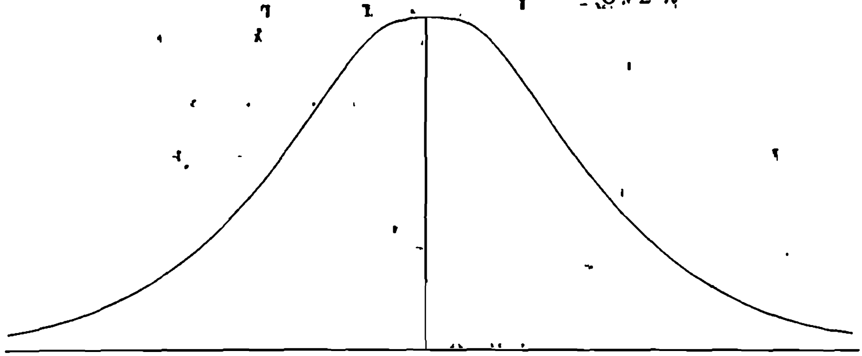
(*vi*) Since $f(x)$ being the probability, can never be negative, no portion of the curve lies below the x-axis.

(*vii*) Linear combination of independent normal variates is also a normal variate.

(*viii*) x-axis is an asymptote to the curve.

(*ix*) The points of inflexion of the curve are given by

$$\left[x = \mu \pm \sigma, f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}\right]$$



$X = \mu$

(*Normal Probability Cuve*)

(*x*) Mean deviation about mean is $\left.\begin{matrix}\\ \sqrt{2/\pi}\,\sigma \approx \frac{4}{5}\sigma \text{ (approx.)}\end{matrix}\right\}$ Q.D. $= \dfrac{Q_3 - Q_1}{2} \simeq \dfrac{2}{3}\sigma$

We have (approximately)

$$Q.D. : M.D. : S.D. :: \frac{2}{3}\sigma : \frac{4}{5}\sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1$$

$\Rightarrow$ $\quad Q.D. : M.D. : S.D. :: 10 : 12 : 15$

(*xi*) *Area Property*

$$P(\mu - \sigma < X < \mu + \sigma) = 0{\cdot}6826$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{\cdot}9544$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0{\cdot}9973$$

The following table gives the area under the normal probability curve for some important values of standard normal variate Z.

Distances from the mean ordinates in terms of $\pm\sigma$	Area under the curve
$Z = \pm 0{\cdot}745$	$50\% = 0{\cdot}50$
$Z = \pm 1{\cdot}00$	$68{\cdot}26\% = 0{\cdot}6826$
$Z = \pm 1{\cdot}96$	$95\% = 0{\cdot}95$
$Z = \pm 2{\cdot}0$	$95{\cdot}44\% = 0{\cdot}9544$
$Z = \pm 2{\cdot}58$	$99\% = 0{\cdot}99$
$Z = \pm 3{\cdot}0$	$99{\cdot}73\% = 0{\cdot}9973$

(*xii*) If X and Y are independent standard normal variates, then it can be easily proved that $U = X + Y$ and $V = X - Y$ are independently distributed, $U \sim N(0, 2)$ and $V \sim N(0, 2)$.

We state (without proof) the converse of this result which is due to D. Bernstein.

Bernstein's Theorem. If X and Y are independent and identically distributed random variables with finite varaince and if $U = X + Y$ and $V = X - Y$ are independent, then all r.v.'s X, Y, U and V are normally distributed.

(*xiii*) We state below another result which characterises the normal distribution.

If $X_1, X_2, ..., X_n$ are *i.i.d. r.v.'s* with finite variance, then the common distribution is normal if and only if :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

or $$\sum_{i=1}^{n} X_i \text{ and } \sum_{i=1}^{n}(X_i - \bar{X})^2$$

are independent. **[For 'If part', see Theorem 13.5]**

In the following sequences we shall establish some of these properties.

8·2·3. Mode of Normal Distribution. Mode is the value of x for which $f(x)$ is maximum, *i.e.*, mode is the solution of,

$$f'(x) = 0 \text{ and } f''(x) < 0$$

For normal distribution with mean μ and standard deviation σ,

$$\log f(x) = c - \frac{1}{2\sigma^2}(x - \mu)^2,$$

where $c = \log(1/\sqrt{2\pi}\,\sigma)$, is a constant.

Differentiating w.r.t. x, we get

$$\frac{1}{f(x)}\cdot f'(x) = -\frac{1}{\sigma^2}(x - \mu) \Rightarrow f'(x) = -\frac{1}{\sigma^2}(x - \mu) f(x)$$

and $$f''(x) = -\frac{1}{\sigma^2}\Big[1 \cdot f(x) + (x-\mu) f'(x)\Big] = -\frac{f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right] \quad \ldots(8\cdot6)$$

Now $f'(x) = 0 \Rightarrow x - \mu = 0$ *i.e.*, $x = \mu$

At the point $x = \mu$, we have from (8·6)

$$f''(x) = -\frac{1}{\sigma^2}\left[f(x)\right]_{x=\mu} = -\frac{1}{\sigma^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} < 0$$

Hence $x = \mu$, is the mode of the normal distribution.

8·2·4. Median of Normal Distribution. If M is the median of the normal distribution, we have

$$\int_{-\infty}^{M} f(x)\,dx = \frac{1}{2} \Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{M} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx$$

$$+ \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\} dx = \frac{1}{2} \quad \ldots(8\cdot7)$$

But $$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left(-z^2/2\right) dz = \frac{1}{2}$$

$\therefore$ From (8·7), we get

$$\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = \frac{1}{2}$$

$$\Rightarrow \quad \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = 0 \quad \Rightarrow \quad \mu = M$$

Hence for the normal distribution, Mean = Median.

Remark. From § 8·2·3 and § 8·2·4, we find that for the normal distribution mean, median and mode coincide. Hence the distribution is *symmetrical.*

8·2·5. M.G.F. of Normal Distribution. The m.g.f. (about origin) is given by

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{t(\mu + \sigma z)\right\} \exp\left(-z^2/2\right) dz, \qquad \left[z = \frac{x-\mu}{\sigma}\right]$$

$$= e^{\mu t} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\tfrac{1}{2}(z^2 - 2t\sigma z)\right\} dz$$

$$= e^{\mu t}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-\tfrac{1}{2}\left\{(z-\sigma t)^2-\sigma^2 t^2\right\}\right]dz$$

$$= e^{\mu t+t^2\sigma^2/2}\times\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left\{-\tfrac{1}{2}(z-\sigma t)^2\right\}dz$$

$$= e^{\mu t+t^2\sigma^2/2}\times\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(-u^2/2)\,du$$

Hence $$M_X(t) = e^{\mu t+t^2\sigma^2/2} \qquad ...(8·8)$$

Remark. *M.G.F. of Standard Normal Variate.* If $X \sim N(\mu, \sigma^2)$, then standard normal variate is given by

$$Z = (X-\mu)/\sigma$$

Now $M_Z(t) = e^{-\mu t/\sigma}\, M_X(t/\sigma) = \exp(-\mu t/\sigma) . \exp\left(\frac{\mu t}{\sigma}+\frac{t^2}{\sigma^2}\cdot\frac{\sigma^2}{2}\right)$

$$= \exp(t^2/2) \qquad ...(8·8\,a)$$

8·2·6. Cumulant Generating Function (c.g.f.) of Normal Distribution. The c.g.f. of normal distribution is given by

$$K_X(t) = \log_e M_X(t) = \log_e\left(e^{\mu t+t^2\sigma^2/2}\right) = \mu t + \frac{t^2\sigma^2}{2}$$

$\therefore$ Mean $= \kappa_1 =$ Coefficient of t in $K_X(t) = \mu$

Variance $= \kappa_2 =$ Coefficient of $\frac{t^2}{2!}$ in $K_X(t) = \sigma^2$

and $\kappa_r =$ Coefficient of $\frac{t^r}{r!}$ in $K_X(t) = 0$; $r = 3, 4 \ldots$

Thus $\mu_3 = \kappa_3 = 0$ and $\mu_4 = \kappa_4 + 3\kappa_2^2 = 3\sigma^4$

Hence $$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \quad\text{and}\quad \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 \qquad ...(8·9)$$

8·2·7. Moments of Normal Distribution. Odd order moments about mean are given by

$$\mu_{2n+1} = \int_{-\infty}^{\infty}(x-\mu)^{2n+1}f(x)\,dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-\mu)^{2n+1}\exp\left\{-(x-\mu)^2/2\sigma^2\right\}dx$$

$$\therefore \quad \mu_{2n+1} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma z)^{2n+1}\exp(-z^2/2)\,dz \qquad \left[z = \frac{x-\mu}{\sigma}\right]$$

$$= \frac{\sigma^{2n+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n+1} \exp(-z^2/2)\, dz = 0, \qquad \ldots(8\cdot10)$$

since the integrand $z^{2n+1} e^{-z^2/2}$ is an odd function of z.

Even order moments about mean are given by

$$\mu_{2n} = \int_{-\infty}^{\infty} (x-\mu)^{2n} f(x)\, dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2n} \exp(-z^2/2)\, dz$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n} \exp(-z^2/2)\, dz$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \cdot 2 \int_{0}^{\infty} z^{2n} \exp(-z^2/2)\, dz$$

(since integrand is an even function of z)

$$\therefore \quad \mu_{2n} = \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_0^{\infty} (2t)^n e^{-t} \frac{dt}{\sqrt{2t}} \qquad \left[\frac{z^2}{2} = t\right]$$

$$= \frac{2^n \cdot \sigma^{2n}}{\sqrt{\pi}} \int_0^{\infty} e^{-t}\, t^{(n+\frac{1}{2})-1}\, dt$$

$$\Rightarrow \quad \mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \cdot \Gamma(n + \tfrac{1}{2})$$

Changing n to $(n-1)$, we get

$$\mu_{2n-2} = \frac{2^{n-1} \cdot \sigma^{2n-2}}{\sqrt{\pi}} \, \Gamma(n - \tfrac{1}{2})$$

$$\therefore \quad \frac{\mu_{2n}}{\mu_{2n-2}} = 2\sigma^2 \cdot \frac{\Gamma(n+\frac{1}{2})}{\Gamma(n-\frac{1}{2})} = 2\sigma^2 (n - \tfrac{1}{2}) \; [\because \Gamma(r) = (r-1)\,\Gamma(r-1)]$$

$$\Rightarrow \quad \mu_{2n} = \sigma^2 (2n-1)\, \mu_{2n-2} \qquad \ldots(8\cdot11)$$

which gives the *recurrence relation* for the moments of normal distribution.

From (8·11), we have

$$\mu_{2n} = [\,(2n-1)\,\sigma^2\,]\,[\,(2n-3)\,\sigma^2\,]\,\mu_{2n-4}$$

$$= [\,(2n-1)\,\sigma^2\,]\,[\,(2n-3)\,\sigma^2\,]\,[\,(2n-5)\,\sigma^2\,]\,\mu_{2n-6}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$= [\,(2n-1)\,\sigma^2\,]\,[\,(2n-3)\,\sigma^2\,]\,[\,(2n-5)\,\sigma^2\,] \ldots (3\,\sigma^2)\,(1\,\sigma^2) \cdot \mu_0$$

$$= 1.3.5 \ldots (2n-1)\,\sigma^{2n} \qquad \ldots(8\cdot12)$$

From (8·10) and (8·12) we conclude that *for the normal distribution all odd order moments about mean vanish and the even order moments about mean are given by (8·12).*

Aliter. The above result can also be obtained quite conveniently as follows:

The m.g.f. (about mean) is given by

$$E[e^{t(X-\mu)}] = e^{-\mu t} E(e^{tX}) = e^{-\mu t} M_X(t)$$

where $M_X(t)$ is the m.g.f. (about origin).

$\therefore$ m.g.f. (about mean) $= e^{-\mu t} e^{\mu t + t^2\sigma^2/2} = e^{t^2\sigma^2/2}$

$$= \left[1 + (t^2\sigma^2/2) + \frac{(t^2\sigma^2/2)^2}{2!} + \frac{(t^2\sigma^2/2)^3}{3!} + \ldots + \frac{(t^2\sigma^2/2)^n}{n!} + \ldots \right] \quad \ldots(8\cdot13)$$

The coefficient of $\frac{t^r}{r!}$ in (8·13) gives μ_r, the rth moment about mean. Since there is no term with odd powers of t in (8·13), all moments of odd order about mean vanish.

i.e., $\quad \mu_{2n+1} = 0\,;\, n = 0, 1, 2, \ldots$

and $\quad \mu_{2n} =$ Coefficient of $\dfrac{t^{2n}}{(2n)!}$ in (8·13) $= \dfrac{\sigma^{2n} \times (2n)!}{2^n\, n!}$

$$= \frac{\sigma^{2n}}{2^n\, n!} \cdot \left[2n(2n-1)(2n-2)(2n-3)\ldots 5.4.3.2.1 \right]$$

$$= \frac{\sigma^{2n}}{2^n . n!} \left[1.3.5\ldots(2n-1) \right] \left[2.4.6 \ldots (2n-2).2n \right]$$

$$= \frac{\sigma^{2n}}{2^n . n!} \left[1.3.5 \ldots (2n-1) \right] 2^n \left[1.2.3 \ldots n \right]$$

$$= 1.3.5\ldots(2n-1)\,\sigma^{2n}$$

Remark. In particular, we have from (8·10) and (8·12),

$$\mu_3 = 0 \text{ and } \mu_2 = 1\,.\,\sigma^2,\ \mu_4 = 1\cdot3\ \sigma^4$$

Hence $\quad \beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = 0$ and $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3\sigma^4}{\sigma^4} = 3,$

the results which have already been obtained in (8·9).

8·2·8. A linear combination of independent normal variates is also a normal variate. Let X_i, $(i = 1, 2, \ldots, n)$ be n independent normal variates with mean μ_i and variance σ_i^2 respectively. Then

$$M_{X_i}(t) = \exp\left\{ \mu_i t + (t^2\sigma_i^2/2) \right\} \quad \ldots(8\cdot14)$$

The m.g.f. of their linear combination $\sum_{i=1}^{n} a_i X_i$, where $a_1, a_2, \ldots, a_n$ are constants, is given by

$$M_{\Sigma a_i X_i}(t) = M_{a_1 X_1 + a_2 X_2 + \ldots + a_n X_n}(t)$$

$$= M_{a_1 X_1}(t) \cdot M_{a_2 X_2}(t) \ldots M_{a_n X_n}(t)$$

($\because$ X_i's are independent)

$$= M_{X_1}(a_1 t) \cdot M_{X_2}(a_2 t) \ldots M_{X_n}(a_n t) \qquad \ldots(8\cdot15)$$

$[\because M_{cX}(t) = M_X(ct)]$

From (8·14), we have

$$M_{X_i}(a_i t) = e^{\mu_i a_i t + t^2 a_i^2 \sigma_i^2/2}$$

$\therefore$ (8·15), gives

$$M_{\sum_i a_i X_i}(t) = [e^{\mu_1 a_1 t + t^2 a_1^2 \sigma_1^2/2} \times e^{\mu_2 a_2 t + t^2 a_2^2 \sigma_2^2/2} \times \ldots \times e^{\mu_n a_n t + t^2 a_n^2 \sigma_n^2/2}]$$

$$= \exp\left[\left(\sum_{i=1}^{n} a_i \mu_i\right) t + t^2 \left(\sum_{i=1}^{n} a_i^2 \sigma_i^2\right) / 2\right],$$

which is the m.g.f. of a normal variate with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i^2 \cdot \sigma_i^2$. Hence by uniqueness theorem of m.g.f.,

$$\sum_{i=1}^{n} a_i X_i \sim N\left[\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right]. \qquad \ldots(8\cdot15\,a)$$

Remarks 1. If we take $a_1 = a_2 = 1$, $a_3 = a_4 = \ldots = 0$, then

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

If we take $a_1 = 1$, $a_2 = -1$, $a_3 = a_4 = \ldots = 0$, then

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

Thus we see that the sum as well as the difference of two independent normal variates is also a normal variate. This result provides a sharp contrast to the Poisson distribution, in which case though the sum of two independent Poisson variates is a Poisson variate, the difference is not a Poisson variate.

2. If we take

$a_1 = a_2 = \ldots = a_n = 1$, then we get $\qquad \ldots(8\cdot15\,b)$

$$\sum_{i=1}^{n} X_i \sim N\left[\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right]$$

i.e., the sum of independent normal variates is also a normal variate, which establishes the *additive property* of the normal distribution.

3. If X_i; $i = 1, 2, \ldots, n$ are identically and independently distributed as $N(\mu, \sigma^2)$ and if we take $a_1 = a_2 = \ldots = a_n = 1/n$,

then
$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left\{\frac{1}{n}\sum_{i=1}^{n} \mu, \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2\right\}$$

$$\Rightarrow \qquad \bar{X} \sim N(\mu, \sigma^2/n), \text{ where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This leads to the following important conclusion :

If X_i, $(i = 1, 2, ..., n)$, are identically and independently distributed normal variates with mean μ and variance σ^2, then their mean $\bar{X}$ is also $N(\mu, \sigma^2/n)$.

8·2·9. Points of Inflexion of Normal Curve. At the point of inflexion of the normal curve, we should have

$$f''(x) = 0, \text{ and } f'''(x) \neq 0$$

For normal curve, we have from (8·6)

$$f''(x) = -\frac{f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right],$$

$$\therefore \qquad f''(x) = 0 \quad \Rightarrow \quad 1 - \frac{(x-\mu)^2}{\sigma^2} = 0 \quad \Rightarrow \quad x = \mu \pm \sigma$$

It can be easily verified that at the points $x = \mu \pm \sigma, f'''(x) \neq 0$.

Hence the points of inflexion of the normal curve are given by $x = \mu \pm \sigma$ and $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}$ *i.e.*, they are equi-distant (at a distance σ) from the mean.

8·2·10. Mean Deviation from the Mean for Normal Distribution.

$$\text{M.D. (about mean)} = \int_{-\infty}^{\infty} |x - \mu| f(x)\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |x - \mu| e^{-(x-\mu)^2/2\sigma^2} dx$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-z^2/2} dz \qquad \left[\frac{x-\mu}{\sigma} = z\right]$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \cdot \int_{0}^{\infty} |z| e^{-z^2/2} dz,$$

since the integrand $|z| e^{-z^2/2}$ is an even function of z.

Since in $[0, \infty]$, $|z| = z$, we have

$$\text{M.D. (about mean)} = \sqrt{2/\pi}\ \sigma \int_{0}^{\infty} z e^{-z^2/2} dz$$

$$= \sqrt{2/\pi}\ \sigma \int_{0}^{\infty} e^{-t} dt \qquad \left[\frac{z^2}{2} = t\right]$$

$$= \sqrt{2/\pi}\ \sigma \left| \frac{e^{-t}}{-1} \right|_0^\infty$$

$$= \sqrt{2/\pi}\ \sigma$$

$$= \frac{4}{5}\sigma \text{ (approx.)}$$

8·2·11. Area Property (Normal Probability Integral). If $X \sim N(\mu, \sigma^2)$, then the probability that random value of X will lie between $X = \mu$ and $X = x_1$ is given by

$$P(\mu < X < x_1) = \int_\mu^{x_1} f(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_\mu^{x_1} e^{-(x-\mu)^2/(2\sigma^2)}\, dx$$

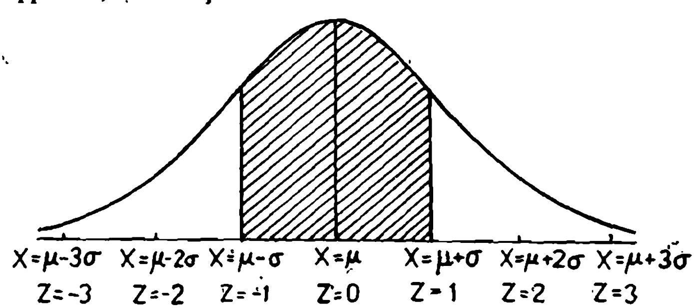
Put $\dfrac{X-\mu}{\sigma} = Z$, *i.e.*, $X - \mu = \sigma Z$

When $X = \mu$, $Z = 0$ and when $X = x_1$, $Z = \dfrac{x_1 - \mu}{\sigma} = z_1$, (say).

$$\therefore \quad P(\mu < X < x_1) = P(0 < Z < z_1) = \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-z^2/2}\, dz = \int_0^{z_1} \varphi(z)\, dz$$

where $\varphi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-z^2/2}$, is the probability function of standard normal variate.

The definite integral $\int_0^{z_1} \varphi(z)\, dz$ is known as *normal probability integral* and gives the area under standard normal curve between the ordinates at $Z = 0$ and $Z = z_1$. These areas have been tabulated for different values of z_1, at intervals of 0·01 [c.f. Appendix, Table IV].



In particular, the probability that a random value of X lies in the interval $(\mu - \sigma, \mu + \sigma)$ is given by

$$P(\mu - \sigma < X < \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)\, dx$$

$$\Rightarrow \qquad P(-1 < Z < 1) = \int_{-1}^{1} \varphi(z)\, dz \qquad \left[z = \frac{x-\mu}{\sigma} \right]$$

$$= 2\int_{0}^{1} \varphi(z)\, dz \qquad \text{(By symmetry)}$$

$$= 2 \times 0{\cdot}3413 = 0{\cdot}6826 \qquad \text{(From tables)} \quad ...(8{\cdot}17)$$

Similarly

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < Z < 2) = \int_{-2}^{2} \varphi(z)\, dz$$

$$= 2\int_{0}^{2} \varphi(z)\, dz = 2 \times 0{\cdot}4772 = 0{\cdot}9544 \quad ...(8{\cdot}18)$$

and

$$P(\sigma - 3\sigma < X < \mu + 3\sigma) = P(-3 < Z < 3) = \int_{-3}^{3} \varphi(z)\, dz$$

$$= 2\int_{0}^{3} \varphi(z)\, dz = 2 \times 0{\cdot}49865 = 0{\cdot}9973 \quad ...(8{\cdot}19)$$

Thus the probability that a normal variate X lies outside the range $\mu \pm 3\sigma$ is given by

$$P(|X-\mu| > 3\sigma) = P(|Z| > 3) = 1 - P(-3 \le Z \le 3) = 0{\cdot}0027$$

Thus in all probability, we should expect a normal variate to lie within the range $\mu \pm 3\sigma$, though theoretically, it may range from $-\infty$ to ∞.

Remarks. 1. The total area under normal probability curve is unity, *i.e.*,

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-\infty}^{\infty} \varphi(z)\, dz = 1$$

2. Since in the normal probability tables, we are given the areas under standard normal curve, in numerical problems we shall deal with the standard normal variate Z rather than the variable X itself.

3. If we want to find area under normal curve, we will somehow or other try to convert the given area to the form $P(0 < Z < z_1)$, since the areas have been given in this form in the tables.

8·2·12. Error Function. If $X \sim N(0, \sigma^2)$, then

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/2\sigma^2}, \quad -\infty < x < \infty$$

If we take $h^2 = \dfrac{1}{2\sigma^2}$ then $f(x) = \dfrac{h}{\sqrt{\pi}}\, e^{-h^2 x^2}$

The probability that a random value of the variate lies in the range $\pm x$ is given by

$$P = \int_{-x}^{x} f(x)\,dx = \frac{h}{\sqrt{\pi}} \int_{-x}^{x} e^{-h^2x^2}\,dx$$

$$= \frac{2h}{\sqrt{\pi}} \int_0^x e^{-h^2x^2}\,dx = \frac{2}{\sqrt{\pi}} \int_0^x e^{-h^2x^2}\,(h\,dx) \qquad ...(*)$$

Taking

$\psi(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-y^2}\,dy$, (*) may be re-written as

$$P = \psi(hx) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-h^2x^2}\,(h\,dx) \qquad ...(**)$$

The function $\psi(y)$, known as the *error function*, is of fundamental importance in the *theory of errors* in Astronomy.

8·2·3. Importance of Normal Distribution. Normal distribution plays a very important role in statistical theory because of the following reasons :

(*i*) Most of the distributions occurring in practice, *e.g.*, Binomial, Poisson, Hypergeometric distributions, etc., can be approximated by normal distribution. Moreover, many of the sampling distributions, *e.g.*, Student's '*t*', Snedecor's F, Chi-square distributions, etc., tend to normality for large samples.

(*ii*) Even if a variable is not normally distributed, it can sometimes be brought to normal form by simple transformation of variable. For example, if the distribution of X is skewed, the distribution of $\sqrt{X}$ might come out to be normal [c.f. Variate Transformations at the end of this Chapter].

(*iii*) If $X \sim N(\mu, \sigma^2)$, then

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0{\cdot}9973$$
$$\Rightarrow \quad P(-3 < Z < 3) = 0{\cdot}9973$$
$$\Rightarrow \quad P(|Z| < 3) = 0{\cdot}9973$$
$$\Rightarrow \quad P(|Z| > 3) = 0{\cdot}0027$$

This property of the normal distribution forms the basis of entire *Large Sample* theory.

(*iv*) Many of the distributions of sample statistic (*e.g.*, the distributions of sample mean, sample variance, etc.) tend to normality for large samples and as such they can best be studied with the help of the normal curves.

(*v*) The entire theory of small sample tests, *viz.*, t, F, χ^2 tests-etc., is based on the fundamental assumption that the parent populations from which the samples have been drawn follow normal distribution.

(*vi*) Theory of normal curves can be applied to the graduation of the curves which are not normal.

(*vii*) Normal distribution finds large applications in Statistical Quality Control in industry for setting control limits.

The following quotation due to Lipman rightly reveals the popularity and importance of normal distribution .

"Every body believes in the law of errors (the normal curve), the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is experimental fact."

W J Youden of the National Bureau of Standards describes the importance of the Normal distribution artistically in the following words :

THE NORMAL
LAW OF ERRORS
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALISATIONS OF NATURAL
PHILOSOPHY IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES.
IN THE PHYSICAL AND SOCIAL SCIENCES
AND IN MEDICINE, AGRICULTURE AND
ENGINEERING. IT IS AN INDISPENSABLE TOOL FOR
THE ANALYSIS AND THE INTERPRETATION OF THE
BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT.

The above presentation, strikingly enough, gives the shape of the normal probability curve.

8·2 14. Fitting of Normal Distribution. In order to fit normal distribution to the given data we first calculate the mean μ, (say), and standard deviation σ, (say), from the given data. Then the normal curve fitted to the given data is given by

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\{-(x-\mu)^2/2\sigma^2\},\ -\infty<x<\infty$$

To calculate the expected normal frequencies we first find the standard normal variates corresponding to the lower limits of each of the class intervals. *i.e.*, we compute $z_i=\dfrac{x_i'-\mu}{\sigma}$, where x_i' is the lower limit of the ith class interval. Then the areas under the normal curve to the left of the ordinate at $z=z_i$, say, $\varphi(z_i)$ are computed from the tables. Finally, the areas for the successive class intervals are obtained by subtraction, *viz.*, $\varphi(z_{i+1})-\varphi(z_i)$, $(i=1,2,\ldots)$ and on multiplying these areas by N, we get the expected normal frequencies.

Example 8·10. *Obtain the equation of the normal curve that may be fitted to the following data :*

Class.	60–65	65–70	70–75	75–80	80–85	85–90	90–95	95–100
Frequency.	3	21	150	335	326	135	26	4

Also obtain the expected normal frequencies

Solution. For the given data, we have

$N = 1000,\ \mu = 79{\cdot}945$ and $\sigma = 5{\cdot}545$

Hence the equation of the normal curve fitted to the given data is

$$f(x) = \frac{1000}{\sqrt{2\pi} \times 5{\cdot}545} \exp\left\{ -\frac{1}{2}\left(\frac{x - 79{\cdot}945}{5{\cdot}545} \right)^2 \right\}$$

Theoretical normal frequencies can be obtained as follows .

class	Lower class boundry (X')	$Z = \frac{X' - \mu}{\sigma}$	$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} dz$	$\Delta \varphi(z) = \psi_{z+1} - \psi_z$	Expected frequency $N \Delta \psi(z)$
Below 60	$-\infty$	$-\infty$	0	0·000112	0·12 ≅ 0
60–65	60	− 3·663	0·000112	0·002914	2·914 ≅ 3
65–70	65	−2·745	0·003026	0·031044	31·044 ≅ 31
70–75	70	−1·826	0·034070	0·147870	147·870 ≅ 148
75–80	75	−0·908	0·181940	0·322050	322·050 ≅ 322
80–85	80	0·010	0·503990	0·319300	319·300 ≅ 319
85–90	85	0·928	0·823290	0·144072	144·072 ≅ 144
90–95	90	1·487	0·967362	0·029792	29·792 ≅ 30
95–100	95	2·675	0·997154	0·002733	2·733 ≅ 3
100 and over	100	3·683	0·999887		
Total					1000

Example 8·11. *For a certain normal distribution, the first moment about 10 is 40 and the fourth moment about 50 is 48. What is the arithmetic mean and standard deviation of the distribution ?*

[Delhi Univ. B.Sc. (Hons. Subs.), 1987; Allahabad Univ. B.Sc. 1990]

Solution. We know that if μ_1' is the first moment about the point $X = A$, then arithmetic mean is given by:

$$\text{Mean} = A + \mu_1'$$

We are given

μ_1' (about the point $X = 10$) $= 40 \quad \Rightarrow \quad \text{Mean} = 10 + 40 = 50$

Also we are given

μ_4' (about the point $X = 50$) $= 48$, *i.e.*, $\mu_4 = 48$ $\qquad (\because \text{Mean} = 50)$

But for a normal distribution with standard deviation σ,

$$\mu_4 = 3\sigma^4 \quad \Rightarrow \quad 3\sigma^4 = 48 \quad i.e., \quad \sigma = 2$$

Example 8·12. *X is normally distributed and the mean of X is 12 and S.D. is 4. (a) Find out the probability of the following :*

(*i*) $X \geq 20$, (*ii*) $X \leq 20$, and (*iii*) $0 \leq X \leq 12$

(*b*) *Find* x', *when* $P(X > x') = 0{\cdot}24$.

(*c*) *Find* x_0' *and* x_1', *when* $P(x_0' < X < x_1') = 0{\cdot}50$ *and* $P(X > x_1') = 0{\cdot}25$

Solution. (*a*) We have $\mu = 12, \sigma = 4$, *i.e.*, $X \sim N(12, 16)$.

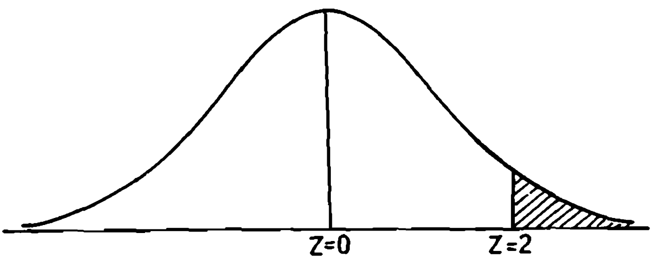
(*i*) $P(X \geq 20) = ?$

When $X = 20$, $\quad Z = \dfrac{20-12}{4} = 2$

$\therefore \quad P(X \geq 20) = P(Z \geq 2) = 0{\cdot}5 - P(0 \leq Z \leq 2) = 0{\cdot}5 - 0{\cdot}4772 = 0{\cdot}0228$

(*ii*) $P(X \leq 20) = 1 - P(X \geq 20)$ $\quad$ ($\because$ Total probability = 1)

$= 1 - 0{\cdot}0228 = 0{\cdot}9772$



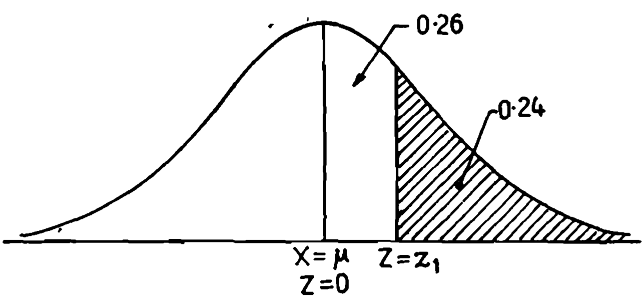
(*iii*) $P(0 \leq X \leq 12) = P(-3 \leq Z \leq 0)$ $\quad \left(Z = \dfrac{X-12}{4}\right)$

$= P(0 \leq Z \leq 3) = 0{\cdot}49865$ $\quad$ (From symmetry)

(*b*) When $X = x'$, $Z = \dfrac{x'-12}{4} = z_1$ (say)

then, we are given

$P(X > x') = 0{\cdot}24 \quad \Rightarrow \quad P(Z > z_1) = 0{\cdot}24, \quad$ *i.e.*, $\quad P(0 < Z < z_1) = 0{\cdot}26$



$\therefore$ From normal tables,

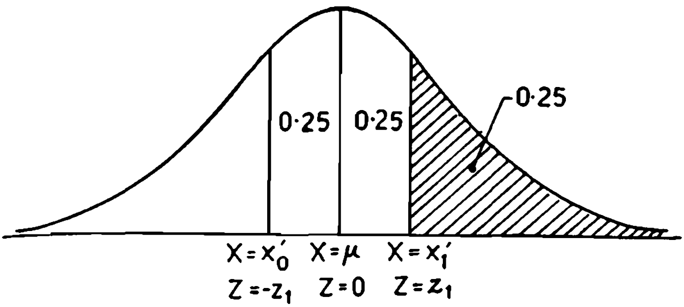
$z_1 = 0{\cdot}71$ (approx.)

Hence $\quad \dfrac{x_1'-12}{4} = 0{\cdot}71 \quad \Rightarrow \quad x_1' = 12 + 4 \times 0{\cdot}71 = 14{\cdot}84$

(*c*) We are given

$P(x_0' < X < x_1') = 0{\cdot}50$ and $P(X > x_1') = 0{\cdot}25$ $\quad$ ...(*)

From (*), obviously the points x_0' and x_1' are located as shown in the figure.



When $X = x_1'$, $\quad Z = \dfrac{x_1' - 12}{4} = z_1$ (say)

and when $X = x_0'$, $\quad Z = \dfrac{x_0' - 12}{4} = -z_1$ (It is obvious from the figure)

We have

$$P(Z > z_1) = 0{\cdot}25 \quad \Rightarrow \quad P(0 < Z < z_1) = 0{\cdot}25$$

$\therefore \quad z_1 = 0{\cdot}67$ (From tables)

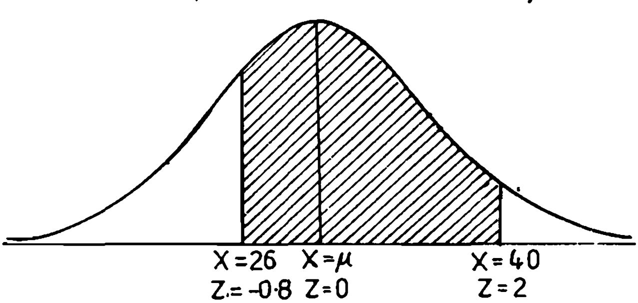
Hence $\quad \dfrac{x_1' - 12}{4} = 0{\cdot}67 \quad \Rightarrow \quad x_1' = 12 + 4 \times 0{\cdot}67 = 14{\cdot}68$

and $\quad \dfrac{x_0' - 12}{4} = -0{\cdot}67 \quad \Rightarrow \quad x_0' = 12 - 4 \times 0{\cdot}67 = 9{\cdot}32$

Example 8·13. *X is a normal variate with mean 30 and S.D. 5. Find the probabilities that*

(*i*) $26 \le X \le 40$, (*ii*) $X \ge 45$, and (*iii*) $|X - 30| > 5$.

Solution. Here $\mu = 30$ and $\sigma = 5$.

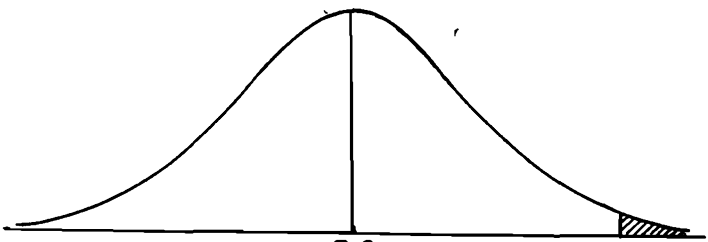


(*i*) When $X = 26$, $Z = \dfrac{X - \mu}{\sigma} = \dfrac{26 - 30}{5} = -0{\cdot}8$

and when $\quad X = 40,\ Z = \dfrac{40 - 30}{5} = 2$

$$\therefore \quad P(26 \le X \le 40) = P(-0{\cdot}8 \le Z \le 2)$$
$$= P(-0{\cdot}8 \le Z \le 0) + P(0 \le Z \le 2)$$
$$= P(-0{\cdot}8 \le Z \le 0) + 0{\cdot}4772 \qquad \text{(From tables)}$$
$$= P(0 \le Z \le 0{\cdot}8) + 0{\cdot}4772 \qquad \text{(From symmetry)}$$
$$= 0{\cdot}2881 + 0{\cdot}4772 = 0{\cdot}7653$$

$$P(X \ge 45) = ?$$



When $X = 45, \quad Z = \dfrac{45 - 30}{5} = 3$

$$\therefore \quad P(X \ge 45) = P(Z \ge 3) = 0{\cdot}5 - P(0 \le Z \le 3)$$
$$= 0{\cdot}5 - 0{\cdot}49865 = 0{\cdot}00135$$

(*iii*) $\quad P(|X - 30| \le 5) = P(25 \le X \le 35) = P(-1 \le Z \le 1)$
$$= 2\,P(0 \le Z \le 1) = 2 \times 0{\cdot}3413 = 0{\cdot}6826$$
$$\therefore \quad P(|X - 30| > 5) = 1 - P(|X - 30| \le 5)$$
$$= 1 - 0{\cdot}6826 = 0{\cdot}3174$$

Example 8·14. *The mean yield for one-acre plot is 662 kilos with a s.d. 32 kilos. Assuming normal distribution, how many one-acre plots in a batch of 1,000 plots would you expect to have yield (i) over 700 kilos, (ii) below 650 kilos, and (iii) what is the lowest yield of the best 100 plots?*

Solution. If the r.v. X denotes the yield (in kilos) for one-acre plot, then we are given that $X \sim N(\mu, \sigma^2)$, where $\mu = 662$ and $\sigma = 32$.

(*i*) The probability that a plot has a yield over 700 kilos is given by

$$P(X > 700) = P(Z > 1{\cdot}19)\,; \qquad Z = \frac{X - 662}{32}$$
$$= 0{\cdot}5 - P(0 \le Z \le 1{\cdot}19)$$
$$= 0{\cdot}5 - 0{\cdot}3830$$
$$= 0{\cdot}1170$$

Hence in a batch of 1,000 plots, the expected number of plots with yield over 700 kilos is $1{,}000 \times 0{\cdot}117 = 117$.

(*ii*) Required number of plots with yield below 650 kilos is given by

$$1000 \times P(X < 650) = 1000 \times P(Z < -0.38) \qquad \left[Z = \frac{650 - 662}{32} \right]$$

$$= 1000 \times P(Z > 0.38) \qquad \text{(By symmetry)}$$

$$= 1000 \times [0.5 - P(0 \le Z \le 0.38)]$$

$$= 1000 \times [0.5 - 0.1480] = 1000 \times 0.352$$

$$= 352$$

(*iii*) The lowest yield, say, x_1 of the best 100 plots is given by

$$P(X > x_1) = \frac{100}{1000} = 0.1$$

When $$X = x_1,\ Z = \frac{x_1 - \mu}{\sigma} = \frac{x_1 - 662}{32} = z_1 \text{ (say)} \qquad \ldots(*)$$

such that $P(Z > z_1) = 0.1 \Rightarrow P(0 \le Z \le z_1) = 0.4$

$\Rightarrow$ $z_1 = 1.28$ (approx.) [From Normal Probability Tables]

Substituting in (*), we get

$$x_1 = 662 + 32\,z = 662 + 32 \times 1.28$$

$$= 662 + 40.96 = 702.96$$

Hence the best 100 plots have yield over 702·96 kilos.

Example 8·15. *There are six hundred Economics students in the post-graduate classes of a university, and the probability for any student to need a copy of a particular book from the university library on any day is 0·05. How many copies of the book should be kept in the university library so that the probability may be greater than 0·90 that none of the students needing a copy from the library has to come back disappointed? (Use normal approximation to the binomial distribution).* **[Delhi Univ. M.A. (Eco.), 1989]**

Solution. We are given :

$$n = 600, p = 0.05, \mu = np = 600 \times 0.05 = 30$$

$$\sigma^2 = npq = 600 \times 0.05 \times 0.95 = 28.5 \Rightarrow \sigma = \sqrt{28.5} = 5.3$$

We want x_1 such that

$$P(X < x_1) > 0.90$$

$$\Rightarrow \quad P(Z < z_1) > 0.90 \qquad \left[z_1 = \frac{x_1 - 30}{5.3} \right]$$

$$\Rightarrow \quad P(0 < Z < z_1) > 0.40$$

$$\Rightarrow \quad z_1 > 1.28 \qquad \text{[From Normal Probability Tables]}$$

$$\Rightarrow \quad \frac{x_1 - 30}{5.3} > 1.28 \quad \Rightarrow \quad x_1 > 30 + 5.3 \times 1.28$$

$$\Rightarrow \quad x_1 > 30 + 6.784 \quad \Rightarrow \quad x_1 > 36.784 \simeq 37$$

Hence the university library should keep at least 37 copies of the book.

Example 8·16. *The marks obtained by a number of students for a certain subject are assumed to be approximately normally distributed with mean value 65* *and with a standard deviation of 5. If 3 students are taken at random from this set what is the probability that exactly 2 of them will have marks over 70?*

Solution. Let the r.v. X denote the marks obtained by the given set of students in the given subject. Then we are given that $X \sim N(\mu, \sigma^2)$ where $\mu = 65$ and $\sigma = 5$. The probability 'p' that a randomly selected student from the given set gets marks over 70 is given by

$$p = P(X > 70)$$

When $\quad X = 70,\ Z = \dfrac{X - \mu}{\sigma} = \dfrac{70 - 65}{5} = 1.$

$$\therefore \quad p = P(X > 70) = P(Z > 1)$$
$$= 0{\cdot}5 - P(0 \le Z \le 1)$$
$$= 0{\cdot}5 - 0{\cdot}3413 = 0{\cdot}1587 \quad \text{[From Normal probability tables]}$$

Since this probability is same for each student of the set, the required probability that out of 3 students selected at random from the set, exactly 2 will have marks over 70, is given by the binomial probability law:

$${}^3C_2\, p^2 \,.\, (1 - p) = 3 \times (0{\cdot}1587)^2 \times (0{\cdot}8413) = 0{\cdot}06357$$

Example 8·17. (*a*) *If $\log_{10} X$ is normally distributed with meant 4 and variance 4, find the probability of*

$$1{\cdot}202 < X < 83180000$$

(Given $\log_{10}$ 1202 = 3·08, $\log_{10}$ 8318 = 3·92).

(b) $\log_{10} X$ is normally distributed with mean 7 and variance 3, $\log_{10} Y$ is normally distributed with mean 3 and variance unity. If the distributions of X and Y are independent, find the probability of $1{\cdot}202 < (X/Y) < 83180000$.

[Given $\log_{10}$ (1202) = 3·08, $\log_{10}$ (8318) = 3·92]

Solution. (*a*) Since $\log X$ is a non-decreasing function of X, we have

$$P(1{\cdot}202 < X < 83180000) = P(\log_{10} 1{\cdot}202 < \log_{10} X < \log_{10} 83180000)$$
$$= P(0{\cdot}08 < \log_{10} X < 7{\cdot}92)$$
$$= P(0{\cdot}08 < Y < 7{\cdot}92)$$

where $Y = \log_{10} X \sim N(4, 4)$ (given).

When $\quad Y = 0{\cdot}08,\ Z = \dfrac{0{\cdot}08 - 4}{2} = -1{\cdot}96$

and when $\quad Y = 7{\cdot}92,\ Z = \dfrac{7{\cdot}92 - 4}{2} = 1{\cdot}96$

$\therefore$ Required probability $= P(0{\cdot}08 < Y < 7{\cdot}92)$

$$= P(-1{\cdot}96 < Z < 1{\cdot}96) = 2P(0 < Z < 1{\cdot}96)$$

(By symmetry)

$$= 2 \times 0{\cdot}4750 = 0{\cdot}9500$$

(*b*) $P[\,1{\cdot}202 < (X/Y) < 83180000\,]$

$$= P[\log_{10} 1{\cdot}202 < \log_{10}(X/Y) < \log_{10} 83180000\,]$$
$$= (0{\cdot}08 < U < 7{\cdot}92)$$

where $\quad U = \log_{10}(X/Y) = \log_{10} X - \log_{10} Y$

Since $\log_{10} X \sim N(7, 3)$ and $\log_{10} Y \sim N(3, 1)$, are independent,

$$\log_{10} X - \log_{10} Y \sim N(7-3, 3+1) \qquad (c.f.\ \text{Remark 1, § 8·2·8})$$

$$\Rightarrow \qquad U = (\log_{10} X - \log_{10} Y) \sim N(4, 4)$$

$\therefore$ Required probability is given by

$$p = P(0{\cdot}08 < U < 7{\cdot}92), \text{ where } U \sim N(4, 4)$$

$$= 0{\cdot}95 \qquad \text{[See part (a)]}$$

Example 8·18. *Two independent random variates X and Y are both normally distributed with means 1 and 2 and standard deviations 3 and 4 respectively. If $Z = X - Y$, write the probability density function of Z. Also state the median, s.d. and mean of the distribution of Z. Find Prob. $\{Z + 1 \le 0\}$.*

Solution. Since $X \sim N(1, 9)$ and $Y \sim N(2, 16)$ are independent, $Z = Y - Y \sim N(1-2, 9+16)$, *i.e.*, $Z = X - Y \sim N(-1, 25)$. Hence p.d.f. of Z is

$$p(z) = \frac{1}{5\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}\left(\frac{z+1}{5}\right)^2\right]; -\infty < z < \infty.$$

For the distribution of Z,

$$\text{Median} = \text{Mean} = -1 \quad \text{and} \quad \text{s.d.} = \sqrt{25} = 5$$

$$P(Z + 1 \le 0) = P(Z \le -1)$$

$$= P(U \le 0); \qquad \left[U = \frac{Z+1}{5} \sim N(0, 1)\right]$$

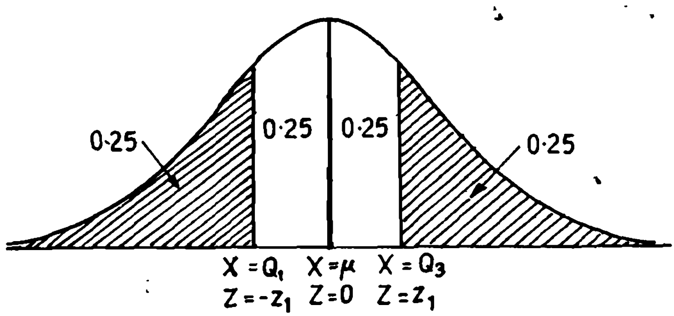
$$= 0{\cdot}5$$

Example 8·19. *Prove that for the normal distribution, the quartile deviation, the mean deviation and standard deviation are approximately 10 : 12 : 15.* **[Dibrugarh Univ. B.Sc. 1993]**

Solution. Let X be a $N(\mu, \sigma^2)$. If Q_1 and Q_3 are the first andthird quartiles respectively, then by definition

$$P(X < Q_1) = 0{\cdot}25 \quad \text{and} \quad P(X > Q_3) = 0{\cdot}25$$

The points Q_1 and Q_3 are located as shown in the figure given below.



When $\quad X = Q_3,\ Z = \dfrac{Q_3 - \mu}{\sigma} = z_1$, (say),

and when $\quad X = Q_1,\ Z = \dfrac{Q_1 - \mu}{\sigma} = -z_1$ (This is obvious from the figure)

Subtracting, we have

$$\frac{Q_3 - Q_1}{\sigma} = 2z_1$$

The quartile deviation is given by

$$Q.D. = \frac{Q_3 - Q_1}{2} = \sigma z_1$$

From the figure, obviously, we have

$P(0 < Z < z_1) = 0{\cdot}25 \Rightarrow z_1 = 0{\cdot}67$ (approx.) (From Normal Tables)

$\therefore \quad Q.D. = \sigma z_1 = 0{\cdot}67\,\sigma = \frac{2}{3}\sigma$

For normal distribution mean deviation about mean (*c.f.* § 8·2·10) is given by

$$\text{M.D.} = \sqrt{2/\pi}\,\sigma = \frac{4}{5}\sigma$$

Hence Q.D. : M.D. : S.D. : : $\frac{2}{3}\sigma : \frac{4}{5}\sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1 :: 10 : 12 : 15$

Example 8·20 (*a*). *In a distribution exactly normal, 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution ?* **[Kerala Univ. B.Sc., May 1991]**

(*b*) *Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming a normal distribution, find the mean height and standard deviation.* **[Nagpur Univ. B.Sc., 1992]**

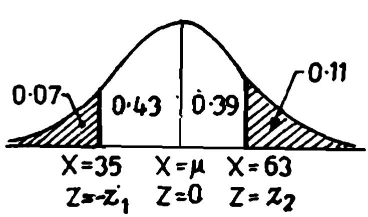
Solution. If $X \sim N(\mu, \sigma^2)$, then we are given

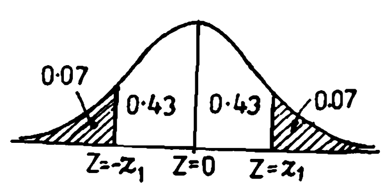
$$P(X < 63) = 0{\cdot}89 \quad \Rightarrow \quad P(X > 63) = 0{\cdot}11 \text{ and } P(X < 35) = 0{\cdot}07$$

The points $X = 63$ and $X = 35$ are located as shown in Fig. (*i*) below.

Since the value $X = 35$ is located to the left of the ordinate at $X = \mu$, the corresponding value of Z is negative.

When $X = 35,\ Z = \dfrac{35 - \mu}{\sigma} = -z_1$, (say),





and when $X = 63$, $Z = \dfrac{63-\mu}{\sigma} = z_2$, (say).

Thus we have, as is obvious from figures (*i*) and (*ii*)

$$P(0 < Z < z_2) = 0{\cdot}39 \text{ and } P(0 < Z < z_1) = 0{\cdot}43$$

Hence from normal tables, we have

$$z_2 = 1{\cdot}23 \text{ and } z_1 = 1{\cdot}48$$

$$\therefore \qquad \frac{63-\mu}{\sigma} = 1{\cdot}23 \text{ and } \frac{35-\mu}{\sigma} = -1{\cdot}48$$

Subtracting, we get

$$\frac{28}{\sigma} = 2{\cdot}71 \Rightarrow \sigma = \frac{28}{2{\cdot}71} = 10{\cdot}33$$

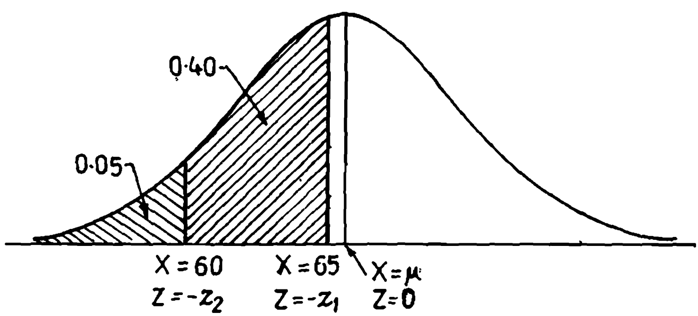
$$\therefore \qquad \mu = 35 + 1{\cdot}48 \times 10{\cdot}33 = 35 + 15{\cdot}3 = 50{\cdot}3$$

(*b*) We are given

$$P(X < 60) = 0{\cdot}05 \text{ and } P(60 < X < 65) = 0{\cdot}40$$

i.e., $\qquad P(X < 65) = 0{\cdot}45$

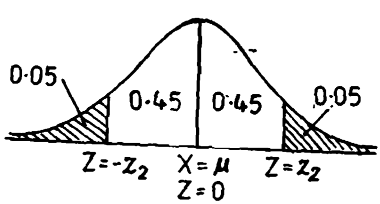
Since the total area to the left of the ordinate at $X = \mu$ is $0{\cdot}5$, both the points $X = 60$ and $X = 65$ are located to the left of $X = \mu$ and consequently the corresponding values of Z are negative.

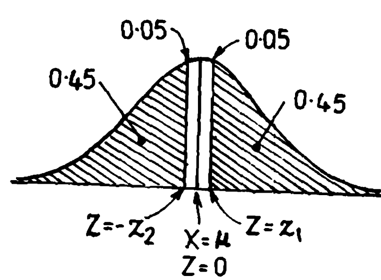


Let $X \sim N(\mu, \sigma^2)$.

When $X = 65$, $\qquad Z = \dfrac{65-\mu}{\sigma} = -z_1$ (say),

and when $X = 60$, $\qquad Z = \dfrac{60-\mu}{\sigma} = -z_2$ (say).





Thus we have

$$P(0 < Z < z_2) = 0.45 \text{ and } P(0 < Z < z_1) = 0.05$$

$\therefore$ $z_2 = 1.645$ and $z_1 = 0.13$ (approx.) (From Normal Tables)

Hence $\dfrac{60-\mu}{\sigma} = -1.645$ ...(*) ; and $\dfrac{65-\mu}{\sigma} = -0.13$ ...(**)

Dividing, we get $\dfrac{60-\mu}{65-\mu} = \dfrac{1.645}{0.13} \Rightarrow \mu = \dfrac{19825}{303} = 65.42$

$\therefore$ From (*), we have $\sigma = \dfrac{60-65.42}{-1.645} = 3.29$

Remarks. If we substitute the value of μ in (**), we get $\sigma = 3.23$ which is only an approximate value since the value of $z_1 = 0.13$, seen from the table, is not exact but only approximate. On the other hand, the value of $z_2 = 1.645$ is exact and hence use of (*) for estimating σ gives better approximation.

Example 8·21 *If the skulls are classified as A, B and C according as the length-breadth index is under 75, between 75 and 80, or over 80, find approximately (assuming that the distribution is normal) the mean and standard deviation of a series in which A are 58%, B are 38% and C are 4%, being given that if*

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t \exp(-x^2/2)\, dx,$$

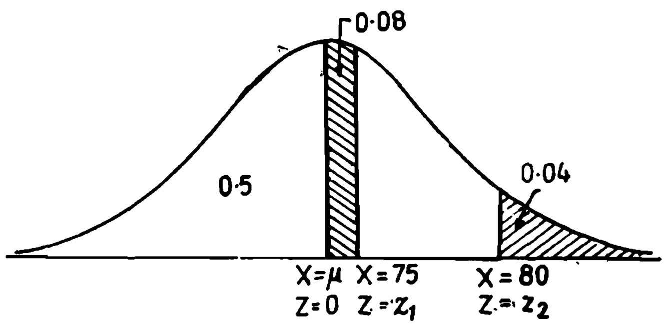
then $f(0.20) = 0.08$ and $f(1.75) = 0.46$

[Delhi Univ. B.Sc., 1989; Burdwan Univ. B.Sc., 1990]

Solution. Let the length-breadth index be denoted by the variable X, then we are given

$$P(X < 75) = 0.58 \text{ and } P(X > 80) = 0.04 \qquad ...(1)$$

Since $P(X < 75)$ represents the total area to the left of the ordinate at the point $X = 75$ and $P(X > 80)$ represents the total area to the right of the ordinate at the point $X = 80$, it is obvious from (1) that the points $X = 75$ and $X = 80$ are located at the positions shown in the figure below.



Now $\frac{1}{\sqrt{2\pi}} \int_0^t \exp(-x^2/2)\, dx$ represents the area under standard normal curve between the ordinates at $Z = 0$ and $Z = t$, Z being a $N(0, 1)$ variate.

$$\therefore \qquad f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t \exp(-x^2/2)\, dx = P(0 < Z < t)$$

Hence $\quad f(0{\cdot}20) = P(0 < Z < 0{\cdot}20) = 0{\cdot}08 \qquad ...(2)$

and $\quad f(1{\cdot}75) = P(0 < Z < 1{\cdot}75) = 0{\cdot}46$

Let μ and σ be the mean and standard deviation of the distribution. Then $X \sim N(\mu, \sigma^2)$.

When $X = 75$, $\qquad Z = \frac{75 - \mu}{\sigma} = z_1$ (say),

and when $X = 80$, $\qquad Z = \frac{80 - \mu}{\sigma} = z_2$ (say).

Thus from the figure, it is obvious that

$$P(X < 75) = 0{\cdot}58 \Rightarrow P(0 < Z < z_1) = 0{\cdot}08$$

$\therefore$ Using (2), we have

$$z_1 = \frac{75 - \mu}{\sigma} = 0{\cdot}20 \qquad ...(3)$$

Also $\qquad P(X > 80) = 0{\cdot}04 \Rightarrow P(0 < Z < z_2) = 0{\cdot}46$

$\therefore$ From (2), we get

$$z_2 = \frac{80 - \mu}{\sigma} = 1{\cdot}75 \qquad ...(4)$$

Solving the equations (3) and (4), we get

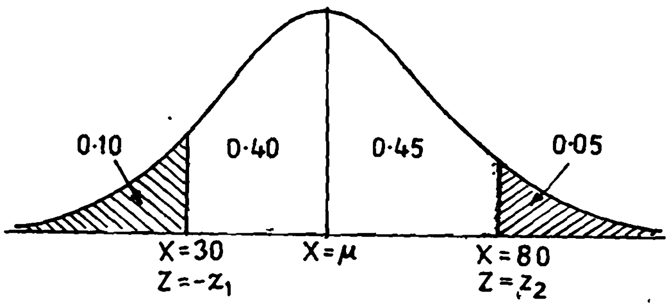
$$\mu = 74{\cdot}4 \text{ (approx.) and } \sigma = 3{\cdot}2 \text{ (approx.)}$$

Example 8·22. *In an examination it is laid down that a student passes if he secures 30 per cent or more marks. He is placed in the first, second or third division according as he secures 60% or more marks, between 45% and 60% marks and marks between 30% and 45% respectively. He gets distinction in case he secures 80% or more marks. It is noticed from the result that 10% of the students failed in the examination, whereas 5% of them obtained distinction. Calculate the percentage of students placed in the second division. (Assume normal distribution of marks.)* **[Aligarh Univ. B.Sc., 1991]**

Solution. Let the variable X denote the marks (out of 100) in the examination and let $X \sim N(\mu, \sigma^2)$. Then we are given

$$P(X < 30) = 0{\cdot}10 \text{ and } P(X \geq 80) = 0{\cdot}05$$

Thus from the figure on next page, we have



When $X = 30$, $\quad Z = \dfrac{30-\mu}{\sigma} = -z_1$ (say),

and when $X = 80$, $\quad Z = \dfrac{80-\mu}{\sigma} = z_2$ (say).

$\therefore \quad P(0 < Z < z_2) = 0{\cdot}5 - 0{\cdot}05 = 0{\cdot}45$

and $\quad P(0 < Z < z_1) = P(-z_1 < Z < 0)$ (By symmetry)

$\quad = 0{\cdot}50 - 0{\cdot}10 = 0{\cdot}40$

$\therefore$ From normal tables, we get

$$z_1 = 1{\cdot}28 \text{ and } z_2 = 1{\cdot}64$$

Hence $\quad \dfrac{30-\mu}{\sigma} = -1{\cdot}28$

$\Rightarrow \quad \dfrac{\mu-30}{\sigma} = 1{\cdot}28$ and $\dfrac{80-\mu}{\sigma} = 1{\cdot}64$

Adding, we get

$$\frac{50}{\sigma} = 2{\cdot}92 \Rightarrow \sigma = \frac{50}{2{\cdot}92} = 17{\cdot}12$$

$\therefore \quad \mu = 30 + 1{\cdot}28 \times 17{\cdot}12 = 30 + 21{\cdot}9136 = 51{\cdot}9136 \simeq 52$

The probability 'p' that a candidate is placed in the second division is equal to the probability that his score lies between 45 and 60, *i.e.*,

$p = P(45 < X < 60) = P(-0{\cdot}41 < Z < 0{\cdot}47)$ $\quad \left[Z = \dfrac{X-52}{17{\cdot}12} \right]$

$\quad = P(-0{\cdot}41 < Z < 0) + P(0 < Z < 0{\cdot}47)$

$\quad = P(0 < Z < 0{\cdot}41) + P(0 < Z < 0{\cdot}47)$ (By symmetry)

$\quad = 0{\cdot}1591 + 0{\cdot}1808 = 0{\cdot}3399 = 0{\cdot}34$ (approx.)

Therefore, 34% candiates got second division in the examination.

Example 8·23. *The local authorities in a certain city instal 10,000 electric lamps in the streets of the city. If these lamps have an average life of 1,000 burning hours with a standard deviation of 200 hours, assuming normality, what number of lamps might be expected to fail (i) in the first 800 burning hours? (ii) between 800 and 1,200 burning hours? After what period of burning hours would you expect that (a) 10% of the lamps would fail? (b) 10% of the lamps would be still burning?*

[In a normal curve, the area between the ordinates corresponding to $\frac{X-\bar{X}}{\sigma}=0$ *and* $\frac{X-\bar{X}}{\sigma}=1$ *is* 0·34134 *and 80% of the area lies between the ordinates corresponding to* $\frac{X-\bar{X}}{\sigma}=\pm 1{\cdot}28$].

Solution. If the variable X denotes the life of a bulb in burning hours, then we are given that $X \sim N(\mu, \sigma^2)$, where $\mu = 1{,}000$ and $\sigma = 200$.

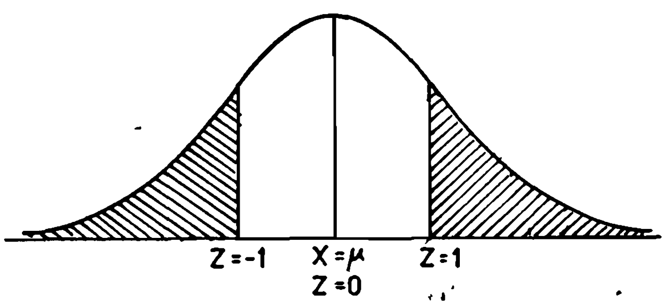
(*i*) The probability 'p' that bulb fails in the first 800 burning hours is given by

$$p = P(X < 800) = P(Z < -1) = P(Z > 1) \qquad \left[Z = \frac{800-1000}{200} \right]$$
$$= 0{\cdot}5 - P(0 < Z < 1) = 0{\cdot}5 - 0{\cdot}3413 = 0{\cdot}1587$$

Therefore out of 10,000 bulbs, the number of bulbs which fail in the first 800 hours is

$$10{,}000 \times 0{\cdot}1587 = 1587$$

(*ii*) Required probability $= P(800 < X < 1200) = P(-1 < Z < 1)$
$$= 2P(0 < Z < 1) = 2 \times 0{\cdot}3413 = 0{\cdot}6826$$



Hence the expected number of blubs with life between 800 and 1,200 hours of burning life is: $10{,}000 \times 0{\cdot}6826 = 6826$

(*a*) Let 10% of the bulbs fail after x_1 hours ot ourning life. Then we have to find x_1 such that $\quad P(X < x_1) = 0{\cdot}10$

When $X = x_1$, $\qquad Z = \frac{x_1 - 1000}{200} = -z_1$ (say).

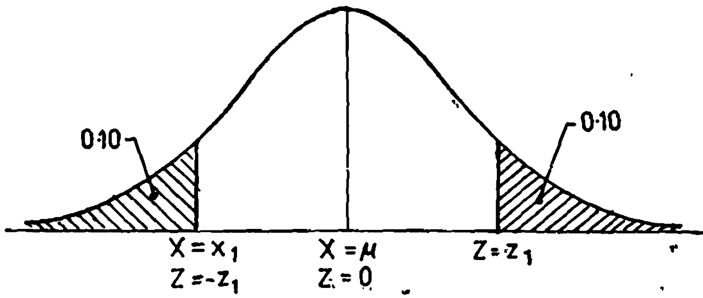
$\therefore \qquad P(Z < -z_1) = 0{\cdot}10 \Rightarrow P(Z > z_1) = 0{\cdot}10$

$\Rightarrow \qquad P(0 < Z < z_1) = 0{\cdot}40 \qquad$ ...(1)

We are given that

$P(-1{\cdot}28 < Z < 1{\cdot}28) = 0{\cdot}80 \Rightarrow 2P(0 < Z < 1{\cdot}28) = 0{\cdot}80$

$\Rightarrow \qquad P(0 < Z < 1{\cdot}28) = 0{\cdot}40 \qquad$ ...(2)



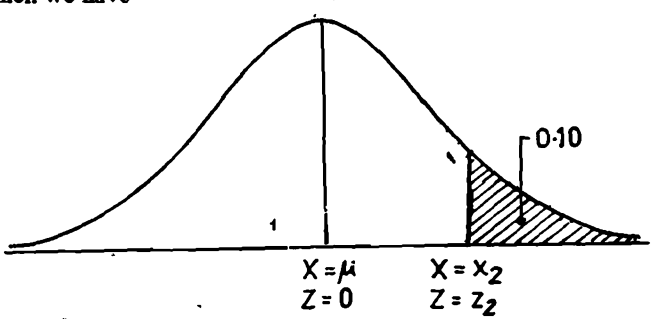
$\therefore$ From (1) and (2), we get

$$z_1 = 1 \cdot 28$$

Hence $\qquad \dfrac{x_1 - 1000}{200} = -1 \cdot 28 \Rightarrow x_1 = 1000 - 256 = 744$

Thus after 744 hours of burning life, 10% of the blubs will fail.

(*b*) Let 10% of the blubs be still burning after, (say), x_2 hours of burning life. Then we have



$P(X > x_2) = 0 \cdot 10 \Rightarrow P(Z > z_2) = 0 \cdot 10 \qquad \left[z_2 = \dfrac{x_2 - 1000}{200} \right]$

$\Rightarrow \qquad P(0 < Z < z_2) = 0 \cdot 40$

$\therefore \qquad z_2 = 1 \cdot 28$ [From (2)]

i.e., $\qquad \dfrac{x_2 - 1000}{200} = 1 \cdot 28 \Rightarrow x_2 = 1000 + 256 = 1256$

Hence after 1256 hours of burning life, 10% of the blubs will be still burning.

Example 8·24. *Let* $X \sim N(\mu, \sigma^2)$. *If* $\sigma^2 = \mu^2, (\mu > 0)$, *express* $P(X < -\mu \mid X < \mu)$ *in terms of cumulative distribution function of* $N(0, 1)$.

[Delhi Univ. B.Sc. (Maths. Hons.) 1988; (Stat. Hons.). 1993]

Solution.

$$P(X < -\mu \mid X < \mu) = \frac{P(X < -\mu \cap X < \mu)}{P(X < \mu)} = \frac{P(X < -\mu)}{P(X < \mu)}; \qquad (\because \mu > 0)$$

$$= \frac{P(Z < -2)}{P(Z < 0)} \qquad \left(Z = \frac{X-\mu}{\sigma} = \frac{X-\mu}{\mu} \right)$$

$$= \frac{P(Z > 2)}{(1/2)}; \qquad \text{(By symmetry)}$$

$$= 2[1 - P(Z \leq 2)] = 2[1 - \Phi(2)]$$

where $\Phi(.)$ is the distribution function of standard normal variate.

Example 8·25 *Can X and $-X$ have the same distribution ? If so, when ?* **[Delhi Univ. B.A., (Spl. Course Statistics), 1989]**

Solution. Yes; X and $-X$ can have the same distribution provided the p.d.f. $f(x)$ of X is symmetric about origin *i.e.,* if $f(-x) = f(x)$.

For example, X and $-X$ have the same distribution if :

(*i*) $X \sim N(0, 1)$

(*ii*) X has standard cauchy distribution [c.f. § 8·9]

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{(1+x^2)}; \quad -\infty < x < \infty$$

(*iii*) X has standard Laplace distribution [c.f. § 8·7]

$$p(x) = \tfrac{1}{2} e^{-|x|}; \quad -\infty < x < \infty.$$

and so on. Obviously X and $Y = -X$ are not identical.

Remark. This example illustrates that if the r.v.'s. X and Y are identical, they have the same distributions. However if X and Y have the same distribution, it does not imply that they are identical.

Example 8·26. *If X, Y are independent normal variates with means 6, 7 and variances 9, 16 respectively, determine λ such that*

$$P(2X + Y \leq \lambda) = P(4X - 3Y \geq 4\lambda)$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1988; B.Sc., 1987]

Solution. Since X and Y are independent, by § 8·2·8 [c.f. equation (8·15*a*)], we have

$$U = 2X + Y \sim N(2 \times 6 + 7, 4 \times 9 + 16), \text{ i.e., } U \sim N(19, 52)$$

$$V = 4X - 3Y \sim N(4 \times 6 - 3 \times 7, 16 \times 9 + 9 \times 16), \text{ i.e., } V \sim N(3, 288)$$

and $\quad P(2X + Y \leq \lambda) = P(U \leq \lambda) = P\left(Z \leq \dfrac{\lambda - 19}{\sqrt{52}} \right)$, where $Z \sim N(0, 1)$

and $\quad P(4X - 3Y \geq 4\lambda) = P(V \geq 4\lambda) = P\left(Z \geq \dfrac{4\lambda - 3}{12\sqrt{2}} \right)$, where $Z \sim N(0, 1)$

Now $\quad P(2X + Y \leq \lambda) = P\{(4X - 3Y) \geq 4\lambda\}$

$$\Rightarrow \quad P\left(Z \leq \frac{\lambda - 19}{\sqrt{52}} \right) = P\left(Z \geq \frac{4\lambda - 3}{12\sqrt{2}} \right)$$

$$\Rightarrow \quad \frac{\lambda - 19}{\sqrt{52}} = -\left(\frac{4\lambda - 3}{12\sqrt{2}} \right)$$

[Since $\quad P(Z \leq a) = P(Z \geq b) \Rightarrow a = -b$,
because normal probability curve is symmetric about $Z = 0$].

$$\Rightarrow \qquad \frac{\lambda - 19}{\sqrt{13}} = \frac{3 - 4\lambda}{6\sqrt{2}}$$

$$\Rightarrow \qquad (6\sqrt{2} + 4\sqrt{13})\,\lambda = 114\sqrt{2} + 3\sqrt{13}$$

$$\Rightarrow \qquad \lambda = \frac{114\sqrt{2} + 3\sqrt{13}}{6\sqrt{2} + 4\sqrt{13}}$$

Example 8·27. *If X and Y are independent normal variates possessing a common mean* μ *such that*

$$P(2X + 4Y \le 10) + P(3X + Y \le 9) = 1$$
$$P(2X - 4Y \le 6) + P(Y - 3X \ge 1) = 1,$$

determine the values of μ and the ratio of the variances of X and Y.

Solution. Let $\text{Var}(X_1) = \sigma_1^2$ and $\text{Var}(Y) = \sigma_2^2$

Since $E(X) = E(Y) = \mu$, (Given) and X and Y are independent by § 8·2·8 [c.f. equation (8·15a)], we have

$2X + 4Y \sim N(2\mu + 4\mu, 4\sigma_1^2 + 16\sigma_2^2)$, *i.e.*, $N(6\mu, 4\sigma_1^2 + 16\sigma_2^2)$

$3X + Y \sim N(3\mu + \mu, 9\sigma_1^2 + \sigma_2^2)$, *i.e.*, $N(4\mu, 9\sigma_1^2 + \sigma_2^2)$

$2X - 4Y \sim N(2\mu - 4\mu, 4\sigma_1^2 + 16\sigma_2^2)$, *i.e.*, $N(-2\mu, 4\sigma_1^2 + 16\sigma_2^2)$

$Y - 3X \sim N(\mu - 3\mu, \sigma_2^2 + 9\sigma_1^2)$, *i.e.*, $N(-2\mu, 9\sigma_1^2 + \sigma_2^2)$

Let us further write :

$$4\sigma_1^2 + 16\sigma_2^2 = \alpha^2 \quad \text{and} \quad 9\sigma_1^2 + \sigma_2^2 = \beta^2 \qquad \text{...(1)}$$

If Z denotes the Standard Normal Variate, *i.e*, if $Z \sim N(0, 1)$, we get

$$P(2X + 4Y \le 10) + P(3X + Y \le 9) = 1$$

$$\Rightarrow \qquad P\left(Z \le \frac{10 - 6\mu}{\alpha}\right) + P\left(Z \le \frac{9 - 4\mu}{\beta}\right) = 1$$

$$\Rightarrow \qquad P\left(Z \le \frac{10 - 6\mu}{\alpha}\right) = 1 - P\left(Z \le \frac{9 - 4\mu}{\beta}\right) = P\left(Z \ge \frac{9 - 4\mu}{\beta}\right)$$

$$\Rightarrow \qquad \frac{10 - 6\mu}{\alpha} = -\left(\frac{9 - 4\mu}{\beta}\right), \qquad \text{...(2)}$$

(Since normal distribution is symmetric about $Z = 0$).

Similarly

$$P(2X - 4Y \le 6) + P(Y - 3X \ge 1) = 1$$

$$\Rightarrow \qquad P\left(Z \le \frac{6 + 2\mu}{\alpha}\right) + P\left(Z \ge \frac{1 + 2\mu}{\beta}\right) = 1$$

$$\Rightarrow \qquad P\left(Z \le \frac{6 + 2\mu}{\alpha}\right) = 1 - P\left(Z \ge \frac{1 + 2\mu}{\beta}\right) = P\left(Z \le \frac{1 + 2\mu}{\beta}\right)$$

$$\Rightarrow \qquad \frac{6 + 2\mu}{\alpha} = \frac{1 + 2\mu}{\beta} \qquad \text{...(3)}$$

Solving (2) and (3), we get

$$\frac{\alpha}{\beta} = \frac{6 + 2\mu}{1 + 2\mu} = \frac{10 - 6\mu}{4\mu - 9} \qquad \text{...(4)}$$

$\Rightarrow \quad (6+2\mu)(4\mu-9)=(10-6\mu)(1+2\mu)$

$\Rightarrow \quad 5\mu^2-2\mu-16=0$ (On simplification)

$$\Rightarrow \quad \mu=\frac{2\pm\sqrt{4+320}}{10}=\frac{2\pm 18}{10}$$

$\Rightarrow \quad \mu=2$ or $-1{\cdot}6$

Substituting $\mu=2$ in (4), we get,

$$\frac{\alpha}{\beta}=\frac{10}{5}=2, \text{ i.e., } 4=\frac{\alpha^2}{\beta^2}$$

From (1), we get

$$4=\frac{4\sigma_1^2+16\sigma_2^2}{9\sigma_1^2+\sigma_2^2}=\frac{4+16\lambda}{9+\lambda} \qquad \left[\text{Taking } \lambda=\frac{\sigma_2^2}{\sigma_1^2}\right]$$

$$\Rightarrow \quad 4(9+\lambda)=4+16\lambda \Rightarrow \lambda=\frac{32}{12}=\frac{8}{3}$$

Again putting $\mu=-1{\cdot}6$ in (4), we get

$$\left(\frac{14}{11}\right)^2=\frac{\alpha^2}{\beta^2}=\frac{4+16\lambda}{9+\lambda} \Rightarrow \lambda=\frac{1280}{1740}=\frac{64}{87}$$

Example 8·28. *If two normal universes A and B have the same total frequency but the standard deviation of universe A is k times that of the universe B, show that maximum frequency of universe A is l/k times that of universe B.*

Solution. Let N be the same total frequency for each of the two universes A and B. If σ is the standard deviation of universe B, then the standard deviation of universe A is $k\sigma$. Let μ_1 and μ_2 be the means of the universes A and B respectively.

The frequency function of universe A is given by

$$f_A(x)=\frac{N}{k\sigma\sqrt{2\pi}}\exp\left\{-(x-\mu_1)^2/2k^2\sigma^2\right\}$$

and the frequency function of universe B is given by

$$f_B(x)=\frac{N}{\sigma\sqrt{2\pi}}\exp\left\{-(x-\mu_2)^2/2\sigma^2\right\}$$

Since, for a normal distribution, the maximum frequency occurs at the point $x=$ mean, we have

$$\begin{aligned}[f_A(x)]_{\max} &= \text{Maximum frequency of universe } A\\ &=\Big[f_A(x)\Big]_{x=\mu_1}\\ &=\left[\frac{N}{k\sigma\sqrt{2\pi}}\exp\left\{-(x-\mu_1)^2/2k^2\sigma^2\right\}\right]_{x=\mu_1}=\frac{N}{k\sigma\sqrt{2\pi}}\end{aligned}$$

Similarly

$$[f_B(x)]_{\max}=[f_B(x)]_{x=\mu_2}$$

$$= \left[\frac{N}{\sigma\sqrt{2\pi}} \exp\left\{ -(x-\mu_2)^2/2\sigma^2 \right\} \right]_{x=\mu_1} = \frac{N}{\sigma\sqrt{2\pi}}$$

$$\therefore \qquad \frac{[f_A(x)]_{max}}{[f_B(x)]_{max}} = \frac{1}{k}$$

EXERCISE 8 (b)

1. "If the Poisson and the Normal distributions are limiting cases of Binomial distribution, then there must be a limiting relation between the Poisson and the Normal distributions." Investigate the relation.

2. (*a*) Derive the mathematical form and properties of normal distribution Discuss the importance of normal distribution in Statistics.

(*b*) Mention the chief characteristics of Normal distribution and Normal probability curve. **[Delhi Univ. B.Sc. (Stat Hons.), 1989]**

3. (*a*) Explain, under what conditions and how the binomial distribution can be approximated to the normal distribution.

(*b*) For a normal distribution with mean 'μ' and standard deviation σ, show that the mean deviation from the mean 'μ' is equal to $\sigma\sqrt{(2/\pi)}$. What will be the mean deviation from median ?

(*c*) The distribution of a variable X is given by the law:

$$f(x) = \text{Constant} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-100}{5} \right)^2 \right], -\infty < x < \infty$$

Write down the value of :

(*i*) the constant,
(*ii*) the mean,
(*iii*) the median,
(*iv*) the mode,
(*v*) standard deviation,
(*vi*) the mean deviation,
(*vii*) the quartile deviation of the distribution.

(Gujarat Univ. B.Sc. April 1978)

Ans. (*i*) $\frac{1}{5\sqrt{2\pi}}$, (*ii*) 100, (*iii*) 100, (*iv*) 100, (*v*) 5 (*vi*) $\sqrt{(2/\pi)} \times 5 \simeq 4$, (*vii*) $\frac{2}{3} \times 5 = 3{\cdot}33$ (approx.).

(*d*) Define Normal probability distribution. If the mean of a Normal population is μ and its variance σ^2, what are its (*i*) mode, (*ii*) Median, (*iii*) β_1 and β_2 ?

(*e*) For a normal distribution $N(\mu, \sigma^2)$:

(*i*) Show that the mean, the median and the mode coincide.
(*ii*) Find the recurrence relation between μ_{2n} and μ_{2n-2}.
(*iii*) State and prove additive property of normal variates.
(*iv*) Obtain the points of inflexion for the normal distribution $N(\mu, \sigma^2)$.
(*v*) Obtain mean deviation about mean.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(*f*) Show that any linear combination of n independent normal variates is also a normal variate. **[Delhi Univ. B.Sc. (Stat. Hons.), 1989]**

(*g*) Show that for the normal curve :

(*i*) The maximum occurs at the mean of the distribution, and

(*ii*) the points of inflexion lie at a distance of $\pm \sigma$ from the mean, where σ is the standard deviation. **[Delhi Univ. M.A. (Eco.), 1987]**

(*h*) Describe the steps involved in fitting a normal distribution to the given data and computing the expected frequencies.

(*i*) Explain how the normal probability integral

$$\int_0^{z_1} \varphi(z)\, dz ,$$

is used in computing normal probabilities.

4. Write a note on the salient features of a normal distribution. $N(\mu, \sigma^2)$ denotes the normal distribution of each of the random variables $X_1, X_2, X_3, ..., X_n$, where μ is the mean and σ^2 the variance. Prove the following :

(*i*) If $X_1, X_2, ..., X_n$ are independent, then $X_1 + X_2 + ... + X_n$ has the distribution $N(n\mu, n\sigma^2)$.

(*ii*) kX, where k is a constant has the distribution $N(k\mu, k^2\sigma^2)$.

(*iii*) $X + a$, where a is a constant has the distribution $N(\mu + a, \sigma^2)$

(*iv*) In (*i*) if $\bar{X} = \dfrac{X_1 + X_2 + ... + X_n}{n}$ then

$$\sqrt{n}\,\frac{(\bar{X} - \mu)}{\sigma} \text{ has the distribution } N(0, 1).$$

5. (*a*) Show that for a normal distribution with mean μ and variance σ^2, the central moments satisfy the relation

$$\mu_{2n} = (2n - 1)\,\mu_{2n-2}\,\sigma^2 \,;\ \mu_{2n+1} = 0$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

Hence show that $\mu_{2n} = \dfrac{(2n)!}{n!}\left(\tfrac{1}{2}\sigma^2\right)^n$ and $\mu_{2n+1} = 0 \,;\ n = 1, 2, ...$

[Delhi Univ. B.Sc. (Stat Hons.) 1985]

(*b*) State the mathematical equation of a normal curve. Discuss its chief features.

(*c*) Find the moment generating function of the normal distribution (m, σ^2), and deduce that

$$\mu_{2n+1} = 0,$$
$$\mu_{2n} = 1 \cdot 3 \cdot 5 \ldots (2n - 1)\,\sigma^{2n},$$

where μ_n denotes the nth central moment.

[Delhi Univ. B.Sc. (Stat. Hons.) 1990,' 82]

(*d*) Show that all central moments of a normal distribution can be expressed in terms of the standard deviation and obtain the expression in the general case.

[Aligarh Univ. B.Sc. 1992]

(*e*) The normal table gives the values of the integral:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}t^2\right) dt$$

for different values of x.

Explain how to use this table to obtain the proportion of observations of a normal variate with mean μ and S.D. σ, which lie above a given value 'a',

(*i*) where $a > \mu$, (*ii*) where $a < \mu$.

6. (*a*) If X_1 and X_2 and two independent normal variates with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, show that the variables U and V where $U = X_1 + X_2$ and $V = X_1 - X_2$, are independent normal variates. Find the means and variances of U and V.

(*b*) If X_1 and X_2 are independent standard normal variates obtain the p.d.f. of $(X_1 - X_2)/\sqrt{2}$.

Ans. $U = (X_1 - X_2)/\sqrt{2} \sim N(0, 1)$

(*c*) Suppose $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$ are independent r.v.'s.

(*i*) Find the joint distribution of $(X_1 + X_2)/\sqrt{2}$ and $(X_1 - X_2)/\sqrt{2}$.

(*ii*) Argue that $2X_1X_2$ and $X_2^2 - X_1^2$ have the same distribution.

Ans. (*i*) $U = (X_1 + X_2)/\sqrt{2}$ and $V = (X_1 - X_2)/\sqrt{2}$ are independent $N(0, 1)$ variates

(*ii*) **Hint.** $X_2^2 - X_1^2 = 2\left[\dfrac{X_2 + X_1}{\sqrt{2}}\right]\left[\dfrac{X_2 - X_1}{\sqrt{2}}\right] = 2\,(U\,V)$

$= 2 \times$ **[Product of two independent SNV 's]**

$2X_1X_2 = 2 \times$ **[Product of two independent SNV 's]**

Hence the result.

7. (*a*) Let X be normally distributed with mean 8 and s.d. 4. Find

(*i*) $P(5 \leq X \leq 10)$, (*ii*) $P(10 \leq X \leq 15)$, (*ii*) $P(X \geq 15)$, (*iv*) $P(X \leq 5)$.

Ans. (*i*) 0·4649 (*ii*) 0·2684 (*iii*) 0·0401 (*vi*) 0·2266.

(*b*) The standard deviation of a certain group of 1,000 high school grades was 11% and the mean grade 78%. Assuming the distribution to be normal, find

(*i*) How many grades were above 90%?

(*ii*) What was the highest grade of the lowest 10?

(*iii*) What was the interquartile range ?

(*iv*) Within what limits did the middle 90% lie?

Ans. (*i*) 138, (*ii*) 52, (*iii*) $Q_1 = 70{\cdot}575$; $Q_3 = 85{\cdot}425$, and (*iv*) 60 % *to* 96·2%

(*c*) If X is normally distributed with mean 2 and variance 1, find $P(|X - 2| < 1)$. **Ans.** 0·6826 [or $\Phi(1) - \Phi(-1)$]

(*d*) If $X \sim N(\mu = 2, \sigma^2 = 2)$, find $P(|X - 1| \leq 2)$ in terms of distribution function of standard normal variate.

Ans. Probability $= P(-1 \leq X \leq 3) = \Phi(1/\sqrt{2}) - \Phi(-3/\sqrt{2})$

(*e*) If $X \sim N(30, 5^2)$ and $Y \sim N(15, 10^2)$, show that

$$P(26 \leq X \leq 40) = P(7 \leq Y \leq 35).$$

Hint. Each Probability $= P(-0.8 \leq Z \leq 2)$ where $Z \sim N(0, 1)$

(*f*) If $X \sim N(30, 5^2)$, find the probabilities of

(*i*) $26 \leq X \leq 40$, (*ii*) $|X - 30| > 5$, (*iii*) $X \geq 42$, (*iv*) $X \leq 28$

[Bihar P.C.S., 1988]

Ans. (*i*) 0·7653, (*ii*) 0·3174, (*iii*) 0·0082, (*iv*) 0·3446

8. (*a*) In a normal population with mean 15·00 and standard deviation 3·5, it is known that 647 observations exceed 16·25. What is the total number of observations in the population ? **(Sri Venkateswara Univ. B.Sc. April 1990)**

Hint. Let $X \sim N(\mu, \sigma^2)$ where $\mu = 15$ and $\sigma = 3{\cdot}5$.

If N is the total number of observations in the population, then we have to find N such that

$$N \times P(X > 16{\cdot}25) = 647$$

(*b*) Assume the mean heights of soldiers to be 68·22 inches with a variance of 10·8 (in.)^2. How many soldiers in a regiment of 1,000 would you expect to be over 6 feet tall? (Given that the area under the standard normal curve between $X = 0$ and $X = 0{\cdot}35$ is 0·1368 and between $X = 0$ and $X = 1{\cdot}15$ is 0·3746).

Ans. 125 **[Osmania Univ. M.A., 1992]**

9. (*a*) If 100 true coins are thrown, how would you obtain an approximation for the probability of getting (*i*) 55 heads, (*ii*) 55 or more heads, using Tables of Area of normal probability function.

(*b*) Prove that Binomial distribution in certain cases becomes normal.

A six faced dice is thrown 720 times. Explain how an approximate value of the probability of the following events can be found out easily. (Finding out the numerical values of these probabilities is not necessary) :

(*i*) 'six' comes for more than 130 times

(*ii*) chance of 'six' lies between 100 and 140.

10. (*a*) The number (X) of items of a certain kind demanded by customers follows the Poisson law with parameter 9. What stock of this item should a retailer keep in order to have a probability of 0·99 of meeting all demands made on him? Use normal approximation to the Poisson law.

(*b*) Show that the probability that the number of heads in 400 throws of a fair coin lies between 180 and 220 is $\simeq 2F(2) - 1$, where $F(x)$ denotes the standard normal distribution function.

11. In an intelligence test administered to 1,000 children, the average score is 42 and standard deviation 24.

(*i*) Find the number of children exceeding the score 60, and (*ii*) Find the number of children with score lying between 20 and 40. (Assume the normal distribution.) **Ans.** (*i*) 227 (*iii*) 289

12. The mean I.Q. (intelligence quotient) of a large number of children of age 14 was 100 and the standard deviation 16. Assuming that the distribution was normal, find

(*i*) What % of the children had I.Q. under 80?

(*ii*) Between what limits the I.Q.'s of the middle 40% of the children lay?

(*iii*) What % of the children had I.Q.'s within the range $\mu \pm 1\cdot96\,\sigma$?

Ans. (*i*) 10·56%, (*ii*) 91·6, 108·4, (*iii*) 0·95

13. (*a*) In a university examination of a particular year, 60% of the students failed when mean of the marks was 50% and s.d. 5%. University decided to relax the conditions of passing by lowering the pass marks, to show its result 70%. Find the minimum marks for a student to pass, supposing the marks to be normally distributed and no change in the performance of students takes place.

Ans. 47·375.

(*b*) The width of a slot on a forging is normally distributed with mean 0·900 inch and standard deviation 0·004 inch. The specifications are 0·900 ± 0·005 inch. What percentage of forgings will be defective ?

Hint. Let X denote the width (in inches) of the slot. We want

$$100 \times P\,(X \text{ lies outside specification limits})$$
$$= 100\,[\,1 - P\,(X \text{ lies within specification limits})\,]$$
$$= 100\,[\,1 - P\,(0\cdot895 < X < 0\cdot905)\,]\,.$$

14. (*a*) The monthly incomes of a group of 10,000 persons were found to be normally distributed with mean Rs. 750 and s.d. Rs. 50. Show that of this group, about 95% had income exceeding Rs. 668 and only 5% had income exceeding Rs 832. What was the lowest income among the richest 100?

Ans. Rs. 866·3 .

(*b*) Given that X is normally distributed with mean 10 and

$$P\,(X > 12) = 0\cdot1587,$$

what is the probability that X will fall in the interval (9, 11)?

Take $\Phi\,(1) = 0\cdot8413$ and $\Phi\,(-\tfrac{1}{2}) = 0\cdot3085$

where $$\Phi\,(x) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{x} \exp\,(-u^2/2)\,du$$

Ans. 0·3830

(*c*) A normal distribution has mean 25 and variance 25. Find

(*i*) the limits which include the middle 50% of the area under the curve, and

(*ii*) the values of x corresponding to the points of inflexion of the curve..

Ans. (*i*) Limits which include the middle 50% of the area under the curve are:

$Q_1 = \mu - 0\cdot6745\,\sigma = 21\cdot7275$; $Q_3 = \mu + 0\cdot6745\,\sigma = 38\cdot2725$

(*ii*) (30, 20)

15. (*a*) In a distribution exactly normal 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution?

Ans. $\mu = 50{\cdot}3,\ \ \sigma = 10{\cdot}33$.

(*b*) In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and variance of the distribution.

Given that area between mean-ordinates and ordinate at any σ distance from mean,

$$Z = \frac{X - \mu}{\sigma} \ : \ 0{\cdot}496 \quad 1{\cdot}405$$

$$\text{Area} \quad : \quad 0{\cdot}19 \quad 0{\cdot}42$$

[Delhi Univ. B.Sc., 1987; Madras Univ. B.Sc., 1990]

Ans. $\mu = 50, \sigma = 10$

16. (*a*) A minimum height is to be prescribed for eligibility to government services such that 60% of the young men will have a fair chance of coming up to that standard. The heights of youngmen are normally distributed with mean 60·6" and s.d. 2·55". Determine the minimum specification.

Ans. 59·9".

Hint. We want x_1 s.t. $P(X > x_1) = 0{\cdot}6$

When $X = x_1,\ Z = \dfrac{x_1 - 60{\cdot}6}{2{\cdot}55} = -z_1$, (say) . ...(*)

[Note the negative sign, which is obvious from the diagram]

Obviously $P(0 < Z < z_1) = 0{\cdot}10 \Rightarrow z_1 = 0{\cdot}254$

Substituting in (*), we get

$$x_1 = 60{\cdot}6 - 2{\cdot}55 \times 0{\cdot}254 = 60{\cdot}6 - 0{\cdot}65 = 59{\cdot}95''$$

(*b*) The height measurements of 600 adult males are arranged in ascending order and it is observed that 180th and 450th entries are 64·2" and 67·8" respectively. Assuming that the sample of heights is drawn from a normal population, estimate the mean and s.d. of the distribution.

Ans. 67·78", 3"

17. (*a*) Marks secured by students in sections I and II of a class are independently normally distributed with means 50 and 60 respectively and variances 10 and 6 respectively. What is the probability that a randomly chosen student from section II scores more marks than a randomly chosen student from section I? What percentage of students are expected to secure first division (*i.e.*, 60 marks or more) in section I? Write down your results in terms of the standard normal distribution fucntion.

Hint. $X \sim N$ (50, 10), $Y \sim N$ (60, 6) are independent r.v.'s. $U = Y - X \sim N$ (10,16). We want $P(Y > X) = P(U > 0)$.

(*b*) In an examination, the mean and standard deviation (s.d.) of marks in Mathematics and Chemistry are given below

	Mean	s.d.
Maths.	45	10
Chem.	50	15

Assuming the marks in the two subjects to be independent normal variates, obtain the probability that a student scores total marks lying between 100 and 130. [Full marks in each subject are 100]. Given that

$$F(0\cdot28) = 0\cdot1103, \ F(1\cdot94) = 0\cdot4738,$$

where
$$F(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp\left(-\tfrac{1}{2}x^2\right) dx\,.$$

[Bhagalpur Univ. B.Sc., 1990]

18. (*a*) One thousand candidates in an examination were grouped into three classes I, II, III in descending order of merits. The numbers in the first two classes were 50 and 350 respectively. The highest and the lowest marks in class II were 60 and 50 respectively. Assuming the distribution to be normal, prove that the average mark is approximately 48·2 and standard deviation, approximately 7·1. The following data may be used:

The area A is measured from the mean zero to any ordinate X.

$\frac{X}{\sigma}$	A	$\frac{X}{\sigma}$	A
0·2	0·079	1·5	0·433
0·3	0·118	1·6	0·445
0·4	0·155	1·7	0·455

(*b*) In an examination marks obtained by the students in Mathematics, Physics and Chemistry are distributed normally about the means 50, 52 and 48 with S.D. 15, 12, 16 respectively. Find the probability of securing total marks of

(*i*) 180 or above, (*ii*) 90 or below.

$$\left[\frac{1}{\sqrt{2\pi}} \int_{1\cdot2}^{\infty} \exp(-z^2/2)\, dz = 0\cdot1942, \ \frac{1}{\sqrt{2\pi}} \int_{2\cdot4}^{\infty} \exp(-z^2/2)\, dz = 0\cdot0224\right]$$

Ans. 0·1942, 0·0224 **[Agra Univ. B.Sc., 1988]**

19. In a certain examination the percentage of passes and distinctions were 46 and 9 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75 respectively. (Assume the distribution of marks to be normal.) (**Ans.** $\mu = 36\cdot4$, $\sigma = 28\cdot2$)

Also determine what would have been the minimum qualifying marks for admission to a re-examination of the failed candidates, had it been desired that the best 25% of them should be given another opportunity of being examined.

Ans. 29.

20. The local authorities in a certain city installed 2,000 electric lamps in a street of the city. If the lamps have an average life of 1,000 burning hours with a S.D. of 200 hours,

(i) What number of the lamps might be expected to fail in the first 700 burning hours,

(ii) After what periods of burning hours would we expect that

(a) 10% of the lamps would have failed, and

(b) 10% of the lamps would be still burning?

Assume that lives of the lamps are normally distributed.

You are given that $F(1\cdot50) = 0\cdot933$, $F(1\cdot28) = \cdot900$,

where $$F(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$

Ans. (i) 134, (ii) (a) 744, (b) 1256. **[Allahabad Univ. B.Sc., 1987]**

21. (a) The quartiles of a normal distribution are 8 and 14 respectively. Estimate the mean and standard deviation.

Ans. $\mu = 11$, $\sigma = 4\cdot4$.

(b) The third decile and the upper quartile of a normal distribution are 56 and 63 respectively. Find the mean and varaince of the distribution.

Ans. $\mu = 59\cdot1$, $\sigma = 5\cdot8$.

22. (a) 5,000 variates are normally distributed with mean 50 and probable error (semi-interquartile range) 13·49. Without using tables, find the values of the quartiles, median, mode standard deviation and mean deviation. Find also the value of the variate for which cumulative frequency is 1250.

[Meerut Univ. B.Sc., 1989]

Ans. $Q_1 \doteq 36\cdot51$ $Q_3 = 63\cdot49$, $\sigma = 20$, M.D. $= 16$, $x_1 = 36\cdot51$.

(b) The following table gives frequencies of occurrence of a varaible X between certain limits :

Variable X	Frequency
Less than 40	30
40 or more but less than 50	33
50 and more	37

The distribution is exactly normal. Find the distribution and also obtain the frequency between $X = 50$ and $X = 60$. **[Kurukshetra Univ. M.A. (Eco.), 1990]**

Ans. Hint. $50 - \mu = 0\cdot33\,\sigma$; $40 - \mu = -0\cdot52\,\sigma$

$$\mu = 46\cdot12,\ \sigma = 11\cdot76$$

$N.P(50 < X < 60) = 100 \times 0\cdot2517 \simeq 25$

23. (a) Suppose that a doorway being constructed is to be used by a class of people whose heights are normally distributed with mean 70" and standard deviation 3". How long may the doorway be without causing more than 25% of the

people to bump their heads? If the height of the doorway is fixed at 76", how many persons out of 5,000 are expected to bump their heads?

[For a normal distributionthe quartile deviation is 0·6745 times standard deviation. For a standard normal distribution $Z = \dfrac{X - \bar{X}}{\sigma}$, the area under the curve between $Z = 0$ and $Z = 2$ is 0 4762.]

(*b*) A normal population has a coefficient of variation 2% and 8% of the population lies above 120. Find the mean and S.D.

Ans. $\mu = 122, \sigma = 2{\cdot}44$

24. Steel rods are manufactured to be 3 inches in diameter but they are acceptable if they are inside the limits 2·99 inches and 3·01 inches. It is observed that 5% are rejected as oversize and 5% are rejected as undersize. Assuming that the diameters are normally distributed, find the standard deviation of the distribution. Hence calculate, what would be the proportion of rejects if the permissible limits were widened to 2·985 inches and 3·015 inches.

[**Hint.** Let X denote the diameter of the rods in inches and let $X \sim N(\mu, \sigma^2)$. Then we are given

$$P(X > 3{\cdot}01) = 0{\cdot}05 \text{ and } P(X < 2{\cdot}99) = 0{\cdot}05$$

$$\Rightarrow \quad \frac{3{\cdot}01 - \mu}{\sigma} = 1{\cdot}65 \quad \text{and} \quad \frac{2{\cdot}99 - \mu}{\sigma} = -1{\cdot}65$$

Solving we get $\mu = 3$ and $\sigma = \dfrac{1}{165}$

The probability that a random value of X lies within the rejection limits is

$$P(2{\cdot}985 < X < 3{\cdot}015) = P(-2{\cdot}475 < Z < 2{\cdot}475) = 2 \times P(0 < Z < 2{\cdot}475)$$
$$= 2 \times 0{\cdot}4933 = 0{\cdot}9866$$

Hence the probability that X lies outside the rejection limits is

$$1 - 0{\cdot}9866 = 0{\cdot}0134$$

Therefore, the proportion of the rejects outside the revised limits is 0·0134, *i.e.*, 1·34%].

25. Derive the moment generating function of a random variable which has a normal distribution with mean μ and variance σ^2. Hence or otherwise prove that a linear combination of independent normal variates is also normally distributed.

An investor has the choice of two of four investments X_1, X_2, X_3, X_4. The profits from these may be assumed to be independetnly distributed, and

the profit from X_1 is $N(2, 1)$,
the profit from X_2 is $N(3, 3)$,
the profit from X_3 is $N(1, \frac{1}{4})$.
the profit from X_4 is $N(2\frac{1}{2}, 4)$.

(Profits are given in £ 1000 per annum).

Which pair should he choose to maximise his probability of making a total annual profit of at least £ 2000? **(London Univ. B.Sc. 1977)**

26. (*a*) State the important properties of the normal distribution and obtain from the tables the inter-quartile range in terms of its mean μ and standard deviation σ.

Find the mean and standard deviation as well as the inter-quartile range of the following data. Compare the inter-quartile range with that obtained from mean and standard deviation on the assumption of normality.

X (central values) ...	0	1	2	3	4	5	6
f (frequency) ...	5	9	15	32	21	10	8

(*b*) The following table gives Baseball throws for a distance by 303 first year high school girls:

Distance in feet	Number of girls	Distance in feet	Number of girls
15 and under 25	1	85 and under 95	44
25 and under 35	2	95 and under 105	31
35 and under 45	7	105 and under 115	27
45 and under 55	25	115 and under 125	11
55 and under 65	33	125 and under 135	4
65 and under 75	53	135 and under 145	1
75 and under 85	64		

(*i*) Fit a normal distribution and find the theoretical frequencies for the classes of the above frequency distribution.

(*ii*) Find the expected number of girls throwing baseballs at a distance exceeding 105 feet on the basis that the data fit a normal distribution.

27. (*a*) The table given below shows the distribution of heights among freshmen in a college :

Height in inches	61	62	63	64	65	66	67	68
Frequency	4	20	23	75	114	186	212	252
Height in inches	69	70	71	72	73	74		
Frequency	218	175	149	46	18	8		

By comparing the proportion of cases lying between $\mu \pm (2/3)\,\sigma$, $\mu \pm \sigma$, $\mu \pm 2\,\sigma$ and $\mu \pm 3\,\sigma$, for this distribution and for a normal curve, state whether the distribution may be considered normal.

(*b*) Fit a normal distribution to the following data of heights in cms of 200 Indian adult males :

Height in (cms)	*Frequency*
144 — 150	3
150 — 156	12
156 — 162	23
162 — 168	52
168 — 174	61
174 — 180	39
180 — 186	10

(*c*) Two hundred and fifty-five metal rods were cut roughly six inches over size. Finally the lengths of the oversize amount were measured exactly and grouped with 1-inch intervals, there being in all 12 groups. The frequency distribution for the 255 lengths was

Central value : x	1	2	3	4	5	6
Frequency : f	2	10	19	25	40	44
x	7	8	9	10	11	12
f	41	28	25	15	5	1

Fit a normal distribution to the data by the method of ordinates and calculate the expected frequencies.

28. (*a*) Let $X \sim N(\mu, \sigma^2)$. Let

$$\Phi(x) = P[X \leq x],$$

calculate the probabilities of the following events in terms of Φ :

(*i*) $\alpha X + \beta \leq t$, where α, β are finite constants.

(*ii*) $-X \geq t$

(*iii*) $|X| > t$ **[Poona Univ. B.E., 1991]**

(*b*) Determine C such that the following function becomes a distribution function:

$$F(x) = C \int_{-\infty}^{x} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] dy$$

29. (*a*) Determine the constant C so that $C.e^{-2x^2+x}, -\infty < x < \infty$, is a density function. If the random variable X has the resulting density function, then find (*i*) the mean of X, (*ii*) the variance of X and (*iii*) $P(X \geq 1/4)$.

Ans. (*i*) 0·25 (*ii*) 0·25 (*iii*) 0·5

(*b*) If $f(x) = k \cdot \exp. \{-(9x^2 - 12x + 13)\}$, is the p.d.f. of a normal distribution (k, being a constant) find the mean and s.d. of the distribution.

(c) If X is a normal variate with p.d.f. $f(x) = 0.03989 \exp(-0.005\,x^2 + 0.5\,x - 12.5)$, express $f(x)$ in standard form and hence find the mean and variance of X. **[M.S. Baroda Univ. B.Sc., 1991]**

(d) Let the probability function of the normal distribution be

$$P(x) = k e^{-1/8\,x^2 + 2x}, \quad -\infty < x < \infty$$

Find k, μ and σ^2. **[Delhi Univ. B.Sc. (Stat. Hons.), 1985]**

(e) X_1, X_2, X_3, X_4 is a random sample from a normal distribution with mean 100 and variance 25 and $\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$.

State the distribution, expected value and variance of each of the following:

(i) $4\bar{X}$, (ii) $X_1 - 2X_2 + X_3 - 3X_4$,

(iii) $\dfrac{1}{25} \sum\limits_{i=1}^{4} \{X_i - 100\}^2$ **[Bangalore Univ. B.Sc., 1989]**

Ans. (b) Mean $= 2/3$, $\sigma = \dfrac{1}{3\sqrt{2}}$

30. If X is a normal variate with mean 50 and s.d. 10, find $P(Y \le 3137)$, where $Y = X^2 + 1$,

$$\left[\frac{1}{\sqrt{2\pi}} \int_0^{0.6} e^{-x^2/2}\, dx = 0.2258 \right]$$ **[Delhi Univ. B.Sc. (Hons.), 1990]**

Hint. Required Probability $= P(X^2 + 1 \le 3137) = P(-56 \le X \le 56)$].

Ans. 0·7258

31. Let X be normally distributed with mean μ and variance σ^2. Suppose σ^2 is some function of μ, say $\sigma^2 = h(\mu)$. Pick $h(.)$ so that $P(X \le 0)$ does not depend on μ for $\mu > 0$.

Ans. $P(X \le 0) = P(Z \le -\mu/\sqrt{h(\mu)}) = P(Z \le -1)$; independent of μ if we take $h(\mu) = \mu^2$.

32. (a) If X is a standard normal variate, find $E\,|X|$ [Ans. $\sqrt{2/\pi} \simeq 4/4$]

(b) X is a random variable normally distributed with mean zero and variance σ^2. Find $E\,|X|$ **[Delhi Univ. B.Sc. (Stat. Hons.) 1990]**

Hint. $E\,|X|$ = Mean Deviation about origin

= M.D. about mean ($\because$ Mean = 0)

Ans. $\sqrt{(2/\pi)}.\ \sigma \approx \dfrac{4}{5}\sigma$

32. (a) X is a normal variate with mean 1 and variance 4, Y is another normal variate independent of X with mean 2 and variance 3. What is the distribution of $X + 2Y$? **[Punjab Univ. B.Sc. (Hons.) 1993]**

Ans. $X + 2Y \sim N(5, 16)$

(*b*) If X is a normal variate with mean 1 and S.D. 0 6, obtain $P[X>0]$, $P[|X-1|\geq 0{\cdot}6]$ and $P[-1{\cdot}8<X<2{\cdot}0]$. What is the distribution of $4X+5$?

34. (*a*) Let X and Y be two independent random variables each with a distribution which is $N(0, 1)$. Find the probability density function of $U=a_1X+a_2Y$, where a_1 and a_2 are constants.

(*b*) Show that if X_1, X_2 are mutually independent normal variates having means μ_1, μ_2 and standard deviations σ_1, σ_2 respectively, then $U=a_1X_1+a_2X_2$ is also normally distributed.

34. (*c*) If X_i, $(i=1, 2, ..., n)$ are independent $N(\mu_i, \sigma_i^2)$ variates, obtain the distribution of $\sum_{i=1}^{n} a_i X_i$

where a_i, $i=1, 2, ..., n$ are constants. Hence deduce the distributions of :

(*i*) X_1+X_2

(*ii*) X_1-X_2

(*iii*) $\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$; if X_i's are i.i.d. $N(\mu, \sigma^2)$.

How do the results in (*i*) and (*ii*) compare with those in Poisson distribution and result in (*iii*) compare with Cauchy distribution ?

[Delhi Univ. B.Sc. (Stat. Hons.), 1991]

Hint. For Cauchy distribution, see Remark 4, § 8·9·1 .

35. (*a*) If X is normal with mean 2 and standard deviation 3, describe the distribution of $Y=\frac{1}{2}X-1$. Explain how you would find $P(Y\geq\frac{3}{2})$ from the tables.

Hint. (*a*) We are given that $X \sim N(\mu, \sigma^2)$ where $\mu=2, \sigma=3$. The distribution of the new variable $Y=aX+b$ is also normal with

$$\left.\begin{aligned} E(Y) &= E(aX+b)=aE(X)+b=a\mu+b \\ \text{and }\ \mathrm{Var}(Y) &= \mathrm{Var}(aX+b)=a^2\,\mathrm{Var}(X)=a^2\sigma^2 \end{aligned}\right\} \quad ...(*)$$

Hence $Y=\frac{1}{2}X-1 \sim N(\mu_1, \sigma_1^2)$, where μ_1 and σ_1^2 are given by (*) with $a=\frac{1}{2}$ and $b=-1$, *i.e.*,

$$\mu_1=\tfrac{1}{2}\,.\,2-1=0\,;\ \sigma_1^2=(\tfrac{1}{2})^2\,.\,9=\tfrac{9}{4}\,.$$

Thus $Y \sim N(\mu_1, \sigma_1^2)$, where $\mu_1=0$, $\sigma_1=\frac{3}{2}$.

$P(Y\geq\frac{3}{2})=P(Z\geq 1)=0{\cdot}5-P(0<Z<1)=0{\cdot}5-0{\cdot}3413=0{\cdot}1587$.

(*b*) If X and Y are independent standard normal variables and if $Z=aX+bY+c$ where a, b and c are contants, what will be the distribution of Z ? What is the mean, median and standard deviation of the distribution of Z ?

Find $P(Z\leq 0{\cdot}1)$ if $a=1, b=-1$ and $c=0$. **(I.I.T. B. Tech. 1992)**

Hint. $Z \sim N(c, a^2+b^2)$

If $a=1, b=-1, c=0$ then $Z=X-Y \sim N(0, 2)$

$\therefore\ P(Z \le 0{\cdot}1) = P\left(U \le \dfrac{1}{14{\cdot}142}\right)$; $\qquad U = \dfrac{Z-0}{\sqrt{2}} \sim N(0, 1)$

36. Let X be a random variable following normal distribution with mean μ and variance σ^2 and let r be a non-negative integer.

If $\mu_r' = E(X^r)$ and if $\mu_{2r} = [E(X-\mu)^{2r}]$, prove that

(i) $\mu'_{r+2} = 2\mu\mu'_{r+1} + (\sigma^2 - \mu^2)\mu_r' + \sigma^3 \dfrac{d\mu_r'}{d\sigma}$

(ii) $\mu_{2r+2} = \sigma^2 \mu_{2r} + \sigma^3 \dfrac{d\mu_{2r}}{d\sigma}$ **[Madras Univ. B.Sc. (Main), Oct. 1989]**

Hint. (i)

$$\frac{d\mu_r'}{d\sigma} = -\int_{-\infty}^{\infty} \frac{x^r}{\sqrt{2\pi}\,\sigma^2} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx$$

$$+ \int_{-\infty}^{\infty} \frac{x^r (x-\mu)^2}{\sqrt{2\pi}\,\sigma^4} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx$$

$$= -\frac{\mu_r'}{\sigma} + \frac{\mu'_{r+2}}{\sigma^3} - \frac{2\mu\mu'_{r+1}}{\sigma^3} + \frac{\mu^2\mu_r'}{\sigma^3}$$

(ii) $$\frac{d\mu_{2r}}{d\sigma} = -\int_{-\infty}^{\infty} \frac{(x-\mu)^{2r}}{\sqrt{2\pi}\,\sigma^2} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx$$

$$+ \int_{-\infty}^{\infty} \frac{(x-\mu)^{2r+2}}{\sqrt{2\pi}\,\sigma^4} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = -\frac{\mu_{2r}}{\sigma} + \frac{\mu_{2r+2}}{\sigma^3}$$

37. Prove that if the independent random variables X and Y have the probability densities,

$$\frac{h}{\sqrt{\pi}} e^{-h^2x^2} \text{ and } \frac{k}{\sqrt{\pi}} e^{-k^2y^2}, -\infty < (x, y) < \infty$$

then the random variable $U = X + Y$ has the probability density,

$$\frac{l}{\sqrt{\pi}} \cdot e^{-l^2u^2}, -\infty < u < \infty$$

where $$\frac{1}{l^2} = \frac{1}{h^2} + \frac{1}{k^2}$$

38. If $\left[\sum_{i=1}^{n} c_i\mu_i\right]^2 = 9 \sum_{i=1}^{n} c_i^2\sigma_i^2$, find $P\left(0 \le Y \le 2 \sum_{i=1}^{n} c_i\mu_i\right)$,

where $Y = \sum_{i=1}^{n} c_i X_i$, X_i being a normal variate with mean μ_i and variance σ_i^2.

(Allahabad Univ. B.Sc., 1988)

Hint.
We know $Y = \sum_{i=1}^{n} c_i X_i \sim N(\mu, \sigma^2)$, where $\mu = \sum_{i=1}^{n} c_i \mu_i$ and $\sigma^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2$

Since $\left(\sum_i c_i \mu_i \right)^2 = 9 \left(\sum_i c_i^2 \sigma_i^2 \right)$, we have $\mu^2 = 9\sigma^2$ or $\dfrac{\mu}{\sigma} = 3$

If we write $Z = \dfrac{Y - \mu}{\sigma}$, then $Z \sim N(0, 1)$.

$$P\left(0 \le Y \le 2 \sum_{i=1}^{n} c_i \mu_i\right) = P(0 \le Y \le 2\mu) = P(-3 \le Z \le 3) = 0{\cdot}9973$$

39. (*a*) Find the mean deviation about mean for the normal distribution $N(\mu, \sigma^2)$.

(*b*) If $X \sim N(\mu, \sigma^2)$, find the mean and variance of

$$Y = \frac{1}{2}\left[(x - \mu)/\sigma \right]^2$$ **[Punjabi Univ. M.A. (Eco.), 1991]**

Ans. $E(Y) = 1/2$, $\text{Var}(Y) = 1/2$

Remark. Also see Example 8·30, on Gamma distribution.

(*c*) Derive normal distribution as a limiting case of binomial distribution, clearly stating the conditions involved. **[Delhi Univ. B.A. (Stat. Hons.), 1981]**

40. If $f(x)$ is the density function for the normal distribution with mean zero and standard deviation σ, then show that

$$\int_{-\infty}^{+\infty} [f(x)]^2 \, dx = \frac{1}{2\sigma\sqrt{\pi}}$$

Hence show that if the normal distribution is grouped in intervals with total frequency N_1, and N_2 is the sum of the squares of the frequencies, an estimate of σ is $\dfrac{N_1^2}{2 N_2 \sqrt{\pi}}$ **(Gujarat Univ. B.Sc., 1992)**

Hint. $\displaystyle\int_{-\infty}^{\infty} [f(x)]^2 \, dx = \int_{-\infty}^{\infty} \left\{ \frac{1}{\sqrt{2\pi}\,\sigma} \exp(-x^2/2\sigma^2) \right\}^2 dx$

$$= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} e^{-x^2/\sigma^2} \, dx = \frac{1}{2\pi\sigma^2} \cdot \frac{\sqrt{\pi}}{(1/\sigma)}$$

$$= \frac{1}{2\sigma\sqrt{\pi}} \qquad \left(\because \int_{-\infty}^{\infty} e^{-a^2x^2} \, dx = \sqrt{\pi}/a \right)$$

$$N_2 = \int_{-\infty}^{\infty} \{ N_1 f(x) \}^2 \, dx = \frac{N_1^2}{2\sqrt{\pi}\,\sigma}$$

41. Obtain the normal distribution as a limiting case of Poisson distribution when the parameter $\lambda \to \infty$.

42. (*a*) If X is $N(0, 1)$, prove that the p.d.f. of $|X|$ is

$$h(x) = \begin{cases} \sqrt{2/\pi}\exp(-x^2/2), & x \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

(*b*) Let $X \sim N(0, 1)$ and $Y \sim N(0, 1)$ be independent random variables. Show that $X + Y$ is independent of $X - Y$.

43. If $X \sim N(\mu, 9^2)$ and $Y \sim N(\mu, 12^2)$ are independent, and if $P(X + 2Y \le 3) = P(2X - Y \ge 4)$, determine μ.

[Calcutta Univ. B.Sc. (Maths Hons.), 1989]

44. If $X \sim N(0, 1)$ and $Y \sim N(0, 1)$, prove that

(*i*) $\text{Var}(\sin X) > \text{Var}(\cos X)$

(*ii*) $E|X - Y| \le \sqrt{8/\pi}$

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1988]

Hint. (*i*) $X \sim N(0, 1) \Rightarrow \varphi_X(t) = E[\cos tX + i \sin tX) = e^{-t^2/2}$.

$\Rightarrow E(\cos tX) = e^{-t^2/2}$ and $E(\sin tX) = 0$.

Taking $t = 1$ and 2, we get:

$$E(\cos X) = e^{-1/2}\,;\ E(\cos 2X) = e^{-2}\,;\ E(\sin X) = E(\sin 2X) = 0\,.$$

$$\text{Var}(\cos X) = E(\cos^2 X) - (E\cos X)^2 = E\left[\frac{1 + \cos 2X}{2}\right] - [E\cos X]^2$$

$$= \frac{1}{2}(1 - e^{-1})^2 \simeq 0.99$$

Similarly $\text{Var}(\sin X) = E\left[\frac{1 - \cos 2X}{2}\right] - [E\sin X]^2 = \frac{1}{2}(1 - e^{-2}) \simeq 0.43$

(*ii*) Use $|X - Y| \le |X| + |Y|$ and $E|X| = E|Y| = \sqrt{2/\pi}$

or $X - Y \sim N(0, \sigma^2 = 2)$; $\quad E|X - Y| = \sqrt{2/\pi}\,\sigma = \sqrt{4/\pi} < \sqrt{(8/\pi)}$

45. Let X and Y be independent $N(0, 1)$ variates. Let $X = R\cos\theta$, $Y = R\sin\theta$. Show that R and θ are independent variates.

[Delhi Univ. B.A. Hons. (Spl. Course Statistics), 1985]

46. If $X \sim N(0, 1)$, find p.d.f. of $|X|$. Hence or otherwise evaluate $E|X|$. **[Delhi Univ. B.Sc. (Maths. Hons.), 1980]**

Hint. Distribution function $G_Y(y)$ of $Y = |X|$ is given by:

$$G_Y(y) = P(Y \le y) = P(|X| \le y) = P(-y \le X \le y)$$
$$= P(X \le y) - P(X \le -y)$$
$$G_Y(y) = F_X(y) - F_X(-y)\,.$$

where $F(\,.\,)$ is the distribution function of X. Differentiating, the p.d.f. of $Y = |X|$ is given by

$$g_Y(y) = f_X(y) + f_X(-y) = 2 f_X(y)$$

$\Rightarrow \quad g_Y(y) = \sqrt{2/\pi}\,.\,e^{-y^2/2}\,;\ y \ge 0 \qquad$ [By symmetry, since $X \sim N(0, 1)$]

8·2·15. The log-normal Distribution. The positive r.v. X is said to have a log-normal distribution if $\log_e X$ is normally distributed.

Let $Y = \log_e X \sim N(\mu, \sigma^2)$.

For $x > 0$,

$F_X(x) = P(X \le x) = P(\log X \le \log x) = P(Y \le \log x)$

(since $\log X$ is monotonic increasing function)

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\log x} \exp\left\{-(y-\mu)^2/2\sigma^2\right\} dy \qquad [\text{since } Y \sim N(\mu, \sigma^2)]$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{x} \exp\left\{-(\log u-\mu)^2/2\sigma^2\right\} \frac{du}{u} \qquad (y = \log u)$$

For $x \le 0$,

$$F_X(x) = P(X \le x) = 0$$

Let us define

$$f_X(u) = \begin{cases} \dfrac{1}{u\sigma\sqrt{2\pi}} \cdot \exp\left\{-(\log u-\mu)^2/2\sigma^2\right\}, u > 0 \\ 0,\ u \le 0 \end{cases} \qquad \ldots(8\cdot18)$$

Then $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$ for every x and hence $f_X(x)$ defined in (8·18) is a p.d.f. of X.

Remark. If $X \sim N(\mu, \sigma^2)$, then $Y = e^X$ is called a log-normal random variable, since its logarithm $\log Y = X$, is a normal r.v.

Moments. The rth moment about origin is given by

$$\mu_r' = E(X^r) = E(e^{rY}) \qquad [\because Y = \log X \Rightarrow X = e^Y]$$

$$= M_Y(r) \qquad (\textit{m.g.f. of } Y, r \textit{ being the parameter})$$

$$= \exp\left\{\mu r + \tfrac{1}{2} r^2 \sigma^2\right\} \qquad [\because Y \sim N(\mu, \sigma^2)] \qquad \ldots(8\cdot19)$$

Remarks. 1. In particular if we take $\mu = \log \alpha$, $\alpha > 0$ *i.e.*, $\log X \sim N(\log \alpha, \sigma^2)$, then

$$\mu_r' = E(X^r) = \exp\left\{r \cdot \log \alpha + \tfrac{1}{2} r^2 \sigma^2\right\} = \alpha^r \cdot \exp\left\{r^2\sigma^2/2\right\} \qquad \ldots(8\cdot19\,a)$$

$$\therefore \text{ mean} = \mu_1' = \alpha\, e^{\sigma^2/2} \text{ and } \mu_2' = \alpha^2 e^{2\sigma^2}$$

$$\mu_2 = \mu_2' - \mu_1'^2 = \alpha^2 e^{\sigma^2}(e^{\sigma^2} - 1)$$

2. It arises in problems of economics, biology, geology, and relibility theory. In particular it arises in the study of dimensions of particles under pulverisation.

3. If $X_1, X_2, \ldots, X_n$ is a set of independently identically distributed random variables such that mean of each $\log X_i$ is μ and its variance is σ^2, then the product $X_1 X_2 \ldots X_n$ is asymptotically distributed according to logarithmic normal distribution and with mean μ and variance $n\sigma^2$

EXERCISE 8(c)

1. (*a*) Let X be a non-negative random variable such that $\log X = Y$, (say), is normally distributed with mean μ and variance σ^2.

(*i*) Write down the probability density function of X. Find $E(X)$ and Var (X).

(*iii*) Find the median and the mode of the distribution of X.

(*b*) If X is a normally distributed with zero mean and variance σ^2 find the density function of $U = e^X$ Locate the mode of the distribution.

2. A random variable X has the probability density function:

$$f(x) = \frac{1}{\beta x \sqrt{2\pi}}, \exp\left[-\frac{1}{2\beta^2}(\log x - \alpha)^2\right], \ x > 0.$$

Find $E(X)$ and Var (X). **[Punjab Univ. M.A. (Eco.), 1991]**

3. A random variate X has the p.d.f.,

$$f(x) = \begin{cases} \dfrac{1}{x\sqrt{2\pi}} e^{-(\log x)^2/2}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Calculate the mean, mode, standard deviation and coefficient of skewness.

Ans. $\sqrt{e}$, $1/e$, $\sqrt{e(e-1)}$, and $(1 - e^{-3/2})/\sqrt{e-1}$

4. The random variable X has mean m and standard deviation s. If $Y = \log X$ is normally distributed with mean M and standard deviation S, prove that

(*i*) $m = \exp\left[M + \frac{1}{2}S^2\right]$, (*ii*) $1 + \dfrac{s^2}{m^2} = e^{S^2}$

5. Given that X_i are independent logarithmic normal variates with parameters μ_i and σ_i; $i = 2, ..., n$, find the sth raw moment of the variable

$$Y = \Pi\,(a_i X_i)\,;\ i = 1, 2, ..., n$$

6. Show that the log-normal distribution is posilively skewed *i.e.*, mean > median > mode.

Ans. Let $Y = \log X \sim N(\mu, \sigma^2)$

$E(X) = e^{\mu + \sigma^2/2}$; Median $= e^{\mu}$; Mode $= e^{\mu - \sigma^2}$

7. If X and Y are two independent log-normal variates, then XY and X/Y are also log-normal variates.

Hint. Let $\log X \sim N(\mu_1, \sigma_1^2)$; $\log Y \sim N(\mu_2, \sigma_2^2)$; $U = XY$ and $V = (X/Y)$.

$$\left.\begin{array}{l} \log U = \log X + \log Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\ \log V = \log X - \log Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \end{array}\right\} \ (\because X \text{ and } Y \text{ are independent})$$

8. If $X \sim N(0, \sigma^2)$, obtain the distribution of e^X Find out the mean of the distribution. **[Delhi Univ. B.Sc. (Stat. Hons.), 1985]**

9. If $X \sim N(\mu, \sigma^2)$, find the p.d.f. of $Y = e^X$, using the result that $E(e^{tX}) = e^{\mu t + t^2 \sigma^2/2}$.

Find the coefficient of variation of Y. **[Delhi Univ. M.A. (Eco.), 1991]**

8·3. Gamma Distribution. *The continuous random variable X which is distributed according to the probability law :*

$$f(x) = \begin{cases} \dfrac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)}\ ; \lambda > 0, 0 < x < \infty \\ 0, \text{ otherwise} \end{cases} \qquad \ldots(8\cdot20)$$

is *known as a Gamma variate with parameter* λ and referred to as a $\gamma(\lambda)$ variate and its distribution is called the Gamma-distribution.

Remarks. 1. The function $f(x)$ defined above represents a probability function, since

$$\int_0^\infty f(x)\, dx = \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{-x} x^{\lambda-1}\, dx = \frac{1}{\Gamma(\lambda)} \cdot \Gamma(\lambda) = 1$$

2. A continuous random variable X having the following p.d.f. is said to have a gamma distribution with two parameters a and λ.

$$\left.\begin{aligned} f(x) &= \frac{a^\lambda}{\Gamma\lambda} e^{-ax} x^{\lambda-1}\ ; a > 0, \lambda > 0\ ;\ 0 < x < \infty \\ &= 0, \text{ otherwise} \end{aligned}\right\} \qquad \ldots(8\cdot20\,a)$$

We write $X \sim \gamma(a, \lambda)$

Taking $a = 1$ in $(8\cdot20\,a)$ we get $(8\cdot20)$. Hence we may write

$X \sim \gamma(\lambda) = (1, \lambda)$.

3. The *cumulative distribution function,* called incomplete gamma function is

$$F_X(x) = \begin{cases} \displaystyle\int_0^x f(u)\, du = \frac{1}{\Gamma\lambda} \int_0^x e^{-u} u^{\lambda-1}\, du, x > 0 \\ 0, \text{ otherwise} \end{cases} \qquad \ldots(8\cdot20\,b)$$

8·3·1. M.G.F. of Gamma Distribution. M.G.F. about origin is given by

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} f(x)\, dx = \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{tx} e^{-x} x^{\lambda-1}\, dx$$

$$= \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{-(1-t)x} x^{\lambda-1}\, dx = \frac{1}{\Gamma(\lambda)} \cdot \frac{\Gamma(\lambda)}{(1-t)^\lambda},\ |t| < 1$$

$$\therefore \quad M_X(t) = (1-t)^{-\lambda},\ |t| < 1 \qquad \ldots(8\cdot21)$$

8·3·2. Cumulant Generating Fucntion of Gamma Distribution. The cumulant generating fucntion $K_X(t)$ is given by

$$K_X(t) = \log M_X(t) = \log(1-t)^{-\lambda} = -\lambda \log(1-t)\ ; |t| < 1$$

$$= \lambda \left[t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \dots \right]$$

$\therefore$ Mean $= \kappa_1 =$ Coefficient of t in $K_X(t) = \lambda$

$$\mu_2 = \kappa_2 = \text{Coefficient of } \frac{t^2}{2!} \text{ in } K_X(t) = \lambda$$

$$\kappa_3 = \text{Coefficient of } \frac{t^3}{3!} \text{ in } K_X(t) = 2\lambda$$

$$\kappa_4 = \text{Coefficient of } \frac{t^4}{4!} \text{ in } K_X(t) = 6\lambda$$

$$\therefore \quad \mu_4 = \kappa_4 + 3\kappa_2^2 = 6\lambda + 3\lambda^2$$

Hence $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{4\lambda^2}{\lambda^3} = \dfrac{4}{\lambda}$ and $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 3 + \dfrac{6}{\lambda}$

Remarks 1. Like Poisson distribution, the mean and variance of the Gamma distribution are also equal. However, Poisson distribution is discrete while Gamma distribution is continuous.

2. Limiting form of Gamma distribution as $\lambda \to \infty$. We know that if $X \sim \gamma(\lambda)$, then $E(X) = \lambda = \mu$, (say), and $\text{Var}(X) = \lambda = \sigma^2$, (say). Then standard gamma variate is given by

$$Z = \frac{X - \mu}{\sigma} = \frac{X - \lambda}{\sqrt{\lambda}}.$$

$$M_Z(t) = e^{-\mu t/\sigma} M_X(t/\sigma) = e^{-\mu t/\sigma}(1 - t/\sigma)^{-\lambda} \qquad \text{[From (8·21)].}$$

$$= e^{-t\lambda/\sqrt{\lambda}} \left(1 - \frac{t}{\sqrt{\lambda}} \right)^{-\lambda}$$

$$\Rightarrow \quad K_Z(t) = -\sqrt{\lambda}\,.\,t - \lambda \log\left(1 - \frac{t}{\sqrt{\lambda}} \right)$$

$$= -\sqrt{\lambda}\,t + \lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \frac{t^3}{3\lambda^{3/2}} + \dots \right)$$

$$= -\sqrt{\lambda}\,t + \sqrt{\lambda}\,t + \frac{t^2}{2} + o(\lambda^{-1/2})$$

where $O(\lambda^{-1/2})$ are terms containing $\frac{1}{2}$ and higher powers of λ in the denominator.

$$\therefore \quad \lim_{\lambda \to \infty} K_Z(t) = \frac{t^2}{2} \quad \Rightarrow \quad \lim_{\lambda \to \infty} M_Z(t) = e^{t^2/2},$$

which is the m.g.f. of a Standard Normal Variate. Hence by uniqueness theorem of m.g.f., Standard Gamma variate tends to Standard Normal Variate as $\lambda \to \infty$. In other words, Gamma distribution tends to Normal distribution for large values of parameter λ.

3. For the two parameter gamma distribution (8·20 *a*), we have

$$M_X(t) = \left(1 - \frac{t}{a}\right)^{-\lambda} \ ; \ t < a . \qquad \ldots(8\cdot21\,a)$$

Proof is left as an exercise to the reader.

$$K_X(t) = -\lambda \log(1 - t/a) \ ; \ t < a$$

$$= \lambda\left[\frac{t}{a} + \frac{1}{2}\left(\frac{t}{a}\right)^2 + \frac{1}{3}\left(\frac{t}{a}\right)^3 + \ldots\right]$$

$\therefore$ $\quad$ Mean $= k_1 = \lambda/a$

Variance $= k_2 = \lambda/a^2 =$ Mean$/a$ $\qquad \ldots(8\cdot21\,b)$

Hence $\quad$ Variance > Mean if $a < 1$,

Variance = Mean if $a = 1$,

and $\quad$ Variance < Mean if $a > 1$.

8·3·3. Additive Property of Gamma Distribution. *The sum of independent Gamma variates is also a Gamma variate. More precisely, if $X_1, X_2, \ldots, X_k$ are independent Gamma variates with parameters $\lambda_1, \lambda_2, \ldots, \lambda_k$ respectively then $X_1 + X_2 + \ldots + X_k$ is also a Gamma variate with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_k$.*

Proof. Since X_i is a $\gamma(\lambda_i)$ variate,

$$M_{X_i}(t) = (1-t)^{-\lambda_i}$$

The m.g.f. of the sum $X_1 + X_2 + \ldots + X_k$ is given by

$$M_{X_1 + X_2 + \ldots + X_k}(t) = M_{X_1}(t)\, M_{X_2}(t) \ldots M_{X_k}(t)$$

(since $X_1, X_2, \ldots, X_k$ are independent)

$$= (1-t)^{-\lambda_1} (1-t)^{-\lambda_2} \ldots (1-t)^{-\lambda_k}$$

$$= (1-t)^{-(\lambda_1 + \lambda_2 + \ldots + \lambda_k)}$$

which is the m.g.f. of a Gamma variate with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_k$. Hence the result follows by the uniqueness theorem of m.g.f.'s.

Remark. If general, if $X_i \sim \gamma(a, \lambda_i)$, $i = 1, 2, \ldots, n$ are independent r.v.'s then $\sum_{i=1}^{n} X_i \sim \gamma\left(a, \sum_{i=1}^{n} \lambda_i\right)$.

8·4. Beta Distribution of First Kind. *The continuous random variable which is distributed according to the probability law*

$$f(x) = \begin{cases} \dfrac{1}{B(\mu, \nu)} \cdot x^{\mu-1}(1-x)^{\nu-1} \ ; \ (\mu, \nu) > 0, \ 0 < x < 1 \\ 0, \ \text{otherwise} \end{cases} \qquad \ldots(8\cdot22)$$

(where $B(\mu, \nu)$ is the Beta function), is known as a Beta variate of the first kind with parameters μ and ν and is referred to as $\beta_1(\mu, \nu)$ variate and its distribution is called Beta distribution of the first kind.

Remarks. 1. The cumulative distribution function, often called the *Incomplete Beta Function, is*

$$F(x)=\begin{cases}0, x<0\\ \displaystyle\int_0^x \frac{1}{B(\mu,\nu)}\,u^{\mu-1}(1-u)^{\nu-1}\,du\ ;\ 0<x<1,\ (\mu,\nu)>0\\ 1, x>1\end{cases} \qquad \ldots(8\cdot22\ a)$$

2. In particular, if we take $\mu = 1$ and $\nu = 1$ in (8·22) we get:

$$f(x)=\frac{1}{\beta(1,1)}=1,\ 0<x<1 \qquad \ldots(8\cdot22\ a)$$

which is the p.d.f. of uniform distribution on [0, 1].

3. If $X\sim\beta_1(\mu,\nu)$, then it can be easily proved that $1-X\sim\beta_1(\nu,\mu)$.

8·4·1. Constants of Beta Distribution of First Kind.

$$\mu_r'=\int_0^1 x^r f(x)\,dx=\frac{1}{B(\mu,\nu)}\int_0^1 x^{\mu+r-1}(1-x)^{\nu-1}\,dx$$

$$=\frac{1}{B(\mu,\nu)}\,B(\mu+r,\nu)=\frac{\Gamma(\mu+r)\,\Gamma(\nu)}{\Gamma(\mu+r+\nu)}\cdot\frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\,\Gamma(\nu)}$$

$$=\frac{\Gamma(\mu+r)\,\Gamma(\mu+\nu)}{\Gamma(\mu+r+\nu)\,\Gamma(\mu)} \qquad \ldots(8\cdot22\ b)$$

In particular

$$\text{Mean}=\mu_1'=\frac{\Gamma(\mu+1)}{\Gamma(\mu+\nu+1)}\cdot\frac{\Gamma(\mu+\nu)}{\Gamma(\mu)}=\frac{\mu\,\Gamma(\mu)\,\Gamma(\mu+\nu)}{(\mu+\nu)\,\Gamma(\mu+\nu)\,\Gamma(\mu)}=\frac{\mu}{\mu+\nu} \qquad \ldots(8\cdot22\ c)$$

$$[\because \Gamma(k)=(k-1)\,\Gamma(k-1)]$$

$$\mu_2'=\frac{\Gamma(\mu+2)\,.\,\Gamma(\mu+\nu)}{\Gamma(\mu+\nu+2)\,\Gamma(\mu)}=\frac{(\mu+1)\,\mu\,\Gamma(\mu)\,\Gamma(\mu+\nu)}{(\mu+\nu+1)\,(\mu+\nu)\,\Gamma(\mu+\nu)\,\Gamma(\mu)}$$

$$=\frac{\mu\,(1+\mu)}{(\mu+\nu)\,(\mu+\nu+1)}$$

Hence

$$\mu_2=\mu_2'-\mu_1'^2=\frac{\mu\,(1+\mu)}{(\mu+\nu)\,(\mu+\nu+1)}-\left(\frac{\mu}{\mu+\nu}\right)^2$$

$$=\frac{\mu}{(\mu+\nu)^2\,(\mu+\nu+1)}\Big[(\mu+\nu)\,(\mu+1)-\mu\,(\mu+\nu+1)\Big]$$

$$=\frac{\mu\,\nu}{(\mu+\nu)^2\,(\mu+\nu+1)} \qquad \ldots(8\cdot22\ d)$$

Similarly, we have

$$\mu_3=\mu_3'-3\,\mu_2'\,\mu_1'+2\,\mu_1'^3=\frac{2\mu\nu\,(\nu-\mu)}{(\mu+\nu)^3\,(\mu+\nu+1)\,(\mu+\nu+2)}$$

and

$$\mu_4=\mu_4'-4\,\mu_3'\,\mu_1'+6\,\mu_2'\,\mu_1'^2-3\,\mu_1'^4$$

$$=\frac{3\,\mu\,\nu\left\{\mu\,\nu\,(\mu+\nu-6)+2\,(\mu+\nu)^2\right\}}{(\mu+\nu)^4\,(\mu+\nu+1)\,(\mu+\nu+2)\,(\mu+\nu+3)}$$

so that

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{4(\nu-\mu)^2(\mu+\nu+1)}{\mu\nu(\mu+\nu+2)^2}$$

and
$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3(\mu+\nu+1)\left[\mu\nu(\mu+\nu-6)+2(\mu+\nu)^2\right]}{\mu\nu(\mu+\nu+2)(\mu+\nu+3)}$$

The harmonic mean H is given by

$$\frac{1}{H} = \int_0^1 \frac{1}{x} f(x)\,dx = \frac{1}{B(\mu,\nu)}\int_0^1 x^{\mu-2}(1-x)^{\nu-1}\,dx$$

$$= \frac{1}{B(\mu,\nu)} B(\mu-1,\nu) = \frac{\Gamma(\mu-1)\Gamma(\nu)}{\Gamma(\mu+\nu-1)}\cdot\frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)}$$

$$= \frac{\Gamma(\mu-1)(\mu+\nu-1)\Gamma(\mu+\nu-1)}{\Gamma(\mu+\nu-1)(\mu-1)\Gamma(\mu-1)} = \frac{\mu+\nu-1}{\mu-1}$$

$$\therefore \quad H = \frac{\mu-1}{\mu+\nu-1} \qquad \ldots(8\cdot22\,e)$$

8·5. Beta Distribution of Second Kind. *The continuous random variable X which is distributed according to the probability law :*

$$f(x) = \begin{cases} \dfrac{1}{B(\mu,\nu)}\cdot\dfrac{x^{\mu-1}}{(1+x)^{\mu+\nu}}\ ; & (\mu,\nu)>0,\ 0<x<\infty \\ 0, \text{ otherwise} \end{cases} \qquad \ldots(8\cdot23)$$

is *known as a Beta variate of second kind with parameters* μ and ν and is denoted as $\beta_2(\mu,\nu)$ variate and its distribution is called Beta distributed of second kind.

Remark. Beta distribution of second kind is transformed to Beta distribution of first kind by the transformation

$$1+x = \frac{1}{y} \quad\Rightarrow\quad y = \frac{1}{1+x} \qquad \ldots(*)$$

Thus if $X \sim \beta_2(\mu,\nu)$, then Y defined in (*) is a $\beta_1(\mu,\nu)$. The proof is left as an exercise to the reader.

8·5·1. Constants of Beta Distribution of Second Kind.

$$\mu_r' = \int_0^\infty x^r f(x)\,dx = \frac{1}{B(\mu,\nu)}\int_0^\infty \frac{x^{\mu+r-1}}{(1+x)^{\mu+\nu}}\,dx$$

$$= \frac{1}{B(\mu,\nu)}\int_0^\infty \frac{x^{(\mu+r)-1}}{(1+x)^{\mu+r+\nu-r}}\,dx$$

$$= \frac{1}{B(\mu,\nu)}\cdot B(\mu+r,\nu-r) \qquad \ldots(1$$

$$= \frac{\Gamma(\mu+r)\Gamma(\nu-r)}{\Gamma(\mu+\nu)}\cdot\frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)} = \frac{\Gamma(\mu+r)\Gamma(\nu-r)}{\Gamma(\mu)\Gamma(\nu)}$$

In particular

$$\mu_1' = \frac{\Gamma(\mu+1)\,\Gamma(\nu-1)}{\Gamma(\mu)\,\Gamma(\nu)} = \frac{\mu\,\Gamma(\mu)\;\Gamma(\nu-1)}{\Gamma(\mu)\,(\nu-1)\,\Gamma(\nu-1)} = \frac{\mu}{\nu-1}$$

$$\mu_2' = \frac{\Gamma(\mu+2)\,\Gamma(\nu-2)}{\Gamma(\mu)\,\Gamma(\nu)} = \frac{(\mu+1)\,\mu\,\Gamma(\mu)\,\Gamma(\nu-2)}{\Gamma(\mu)\,(\nu-1)\,(\nu-2)\,\Gamma(\nu-2)} = \frac{\mu\,(\mu+1)}{(\nu-1)\,(\nu-2)}$$

$$\mu_2 = \mu_2' - \mu_1'^2 = \frac{\mu\,(\mu+1)}{(\nu-1)\,(\nu-2)} - \left(\frac{\mu}{\nu-1}\right)^2$$

$$= \frac{\mu}{\nu-1}\left[\frac{(\nu-1)\,(\mu+1) - \mu\,(\nu-2)}{(\nu-1)\,(\nu-2)}\right] = \frac{\mu\,(\mu+\nu-1)}{(\nu-1)^2\,(\nu-2)}$$

The harmonic mean H is given by

$$\frac{1}{H} = E\left(\frac{1}{X}\right) = \int_0^\infty \frac{1}{x}\,.f(x)\,dx = \frac{1}{B(\mu,\nu)}\int_0^\infty \frac{x^{\mu-2}}{(1+x)^{\mu+\nu}}\,dx$$

$$= \frac{1}{B(\mu,\nu)}\int_0^\infty \frac{x^{\mu-1-1}}{(1+x)^{\mu-1+\nu+1}}\,dx = \frac{1}{B(\mu,\nu)}\,.\,\mathbf{B(\mu-1,\nu+1),\ \mu>1}$$

$$= \frac{\Gamma(\mu-1)\,\Gamma(\nu+1)}{\Gamma(\mu+\nu)}\cdot\frac{\Gamma(\mu+\nu)}{\Gamma(\nu)\,\Gamma(\nu)} = \frac{\Gamma(\mu-1)\,\nu\,\Gamma(\nu)}{(\mu-1)\,\Gamma(\mu-1)\,\Gamma(\nu)} = \frac{\nu}{\mu-1}$$

Hence $\qquad H = \dfrac{\mu-1}{\nu}$

Example 8·29. *The daily consumption of milk in a city, in excess of 20,000 gallons, is approximately distributed as a Gamma variate with the parameters* $\nu = 2$ and $\lambda = \dfrac{1}{10{,}000}$. *The city has a daily stock of 30,000 gallons. What is the probability that the stock is insufficient on a particular day?*

[Madras Univ. B.Sc. (Stat. Main), 1990]

Solution. If the r.v. X denotes the daily consumption of milk (in gallons) in a city, then the r.v. $Y = X - 20{,}000$ has a **gamma distribution with p.d.f.**

$$g(y) = \frac{1}{(10{,}000)^2\,\Gamma(2)}\,y^{2-1}\,e^{-y/10{,}000} = \frac{y\,e^{-y/10{,}000}}{(10{,}000)^2}\ ;\ 0<y<\infty$$

[See 8·20 (*a*)]

Since the daily stock of the city is 30,000 gallons, the required probability 'p' that the stock is insufficient on a particular day is given by

$$p = P(X > 30{,}000) = P(Y > 10{,}000)$$

$$= \int_{10{,}000}^\infty g(y)\,dy = \int_{10{,}000}^\infty \frac{y e^{-y/10{,}000}}{(10{,}000)^2}\,dy$$

$$= \int_1^\infty z\,e^{-z}\,dz \qquad \text{[Taking } z = y/10{,}000\text{]}$$

Integrating by parts, we get

$$p = \left| -z e^{-z} \right|_1^{\infty} + \int_1^{\infty} e^{-z} dz = e^{-1} - \left| e^{-z} \right|_1^{\infty}$$

$$= e^{-1} + e^{-1} = 2/e$$

Remark. Since $\nu = 2$, the integration is easily done. However, for general values of λ and ν, the integral is evaluated by using tables of Incomplete Gamma Integral, [see Tables of Incomplete Gamma Fucntions, K. Pearson; Cambridge University Press] of the form

$$\int_0^{\alpha} \frac{e^{-x} . x^{n-1}}{\Gamma n} . dx,$$

which have been tabulated for different values of α and n.

Example 8·30. *If $X \sim N(\mu, \sigma^2)$, obtain the p.d.f. of:*

$$U = \tfrac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$$

Solution. Since $X \sim N(\mu, \sigma^2)$,

$$dP(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2} dx, \ -\infty < x < \infty$$

Let $\quad u = \dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2 \quad \Rightarrow \quad \dfrac{x-\mu}{\sigma} = \sqrt{2u}$

$\therefore \quad dx = \dfrac{\sigma}{\sqrt{2u}} du$

Hence probability differential of U is

$$dG(u) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-u} . \frac{\sigma}{\sqrt{2u}} du = \frac{1}{2\sqrt{\pi}} e^{-u} u^{-1/2} du$$

$$= \frac{1}{\sqrt{\pi}} e^{-u} u^{-1/2} du, \ 0 < u < \infty$$

the factor $\frac{1}{2}$ disappearing from the fact that total probability is unity.

$$\therefore \quad dG(u) = \frac{1}{\Gamma(\frac{1}{2})} e^{-u} u^{(1/2)-1} du, \ 0 < u < \infty \qquad [\because \Gamma(\tfrac{1}{2}) = \sqrt{\pi}\,]$$

Hence $\quad U = \dfrac{1}{2}\left(\dfrac{X-\mu}{\sigma}\right)^2$ is a $\gamma(\frac{1}{2})$ variate.

Example 8·31. *Show that the mean value of positive square root of a $\gamma(\mu)$ variate is $\Gamma(\mu+\frac{1}{2})/\Gamma(\mu)$. Hence prove that the mean deviation of a normal variate from its mean is $\sqrt{2/\pi}\ \sigma$, where σ is the standard deviation of the distribution.* **[Delhi Univ. B.Sc. (Stat. Hons.), 1986]**

Solution. Let X be a $\gamma(\mu)$ variate. Then

$$f(x) = \frac{e^{-x} x^{\mu-1}}{\Gamma(\mu)}; \ \mu > 0, 0 < x < \infty$$

$$\therefore \quad E(\sqrt{X}) = \int_0^\infty x^{1/2} f(x)\, dx = \frac{1}{\Gamma(\mu)} \int_0^\infty e^{-x} x^{\mu + (1/2) - 1}\, dx = \frac{\Gamma(\mu + \frac{1}{2})}{\Gamma(\mu)}$$

If $X \sim N(\mu, \sigma^2)$, then

$$U = \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2 \text{ is a } \gamma(\tfrac{1}{2}) \text{ variate.} \qquad (\textit{c.f.} \text{ Example 8·30})$$

$\therefore \quad |X - \mu| = \sqrt{2}\, \sigma\, U^{1/2}$, where U is a $\gamma(\frac{1}{2})$ variate.

Hence mean deviation about mean is given by

$$E|X-\mu| = E(\sqrt{2}\, \sigma\, U^{1/2}) = \sqrt{2}\, \sigma\, E(U^{1/2})$$

$$= \sqrt{2}\, \sigma \frac{\Gamma(\frac{1}{2}+\frac{1}{2})}{\Gamma(\frac{1}{2})} = \frac{\sqrt{2}\, \sigma}{\sqrt{\pi}} = \sqrt{2/\pi}\, \sigma$$

Example 8·32. *If X and Y are independent Gamma variates with parameters* μ *and* ν *respectively, show that the variables*

$$U = X + Y, Z = \frac{X}{X+Y}$$

are independent and that U is a $\gamma(\mu+\nu)$ *variate and Z is a* $\beta_1(\mu, \nu)$ *variate.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1991]

Solution. Since X is a $\gamma(\mu)$ variate and Y is a $\gamma(\nu)$ variate, we have

$$f_1(x)\, dx = \frac{1}{\Gamma(\mu)} e^{-x} x^{\mu-1}\, dx\,;\, 0 < x < \infty, \mu > 0$$

$$f_2(y)\, dy = \frac{1}{\Gamma(\nu)} e^{-y} y^{\nu-1}\, dy\,;\, 0 < y < \infty, \nu > 0$$

Since X and Y are independently distributed, their joint probability differential is given by the compound probability theorem as

$$dF(x,y) = f_1(x) f_2(y)\, dx\, dy = \frac{1}{\Gamma\mu\, \Gamma\nu} e^{-(x+y)} x^{\mu-1} y^{\nu-1}\, dx\ dy$$

Now $u = x + y,\, z = \dfrac{x}{x+y}$, so that $x = uz,\, y = u - x = u(1-z)$

Jacobian of transformation J is given by

$$J = \frac{\partial(x,y)}{\partial(u,z)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} \\ \dfrac{\partial x}{\partial z} & \dfrac{\partial y}{\partial z} \end{vmatrix} = \begin{vmatrix} z & 1-z \\ u & -u \end{vmatrix} = -u$$

As X and Y range from 0 to ∞, u ranges from 0 to ∞ and z from 0 to 1. Hence the joint distribution of U and Z is

$$d\,G(u,z) = g(u,z)\, du\, dz = \frac{1}{\Gamma(\mu)\, \Gamma(\nu)} e^{-u} (uz)^{\mu-1} [u(1-z)]^{\nu-1} |J|\, du\, dz$$

$$= \frac{1}{\Gamma(\mu)\, \Gamma(\nu)} \cdot e^{-u} u^{\mu+\nu-1} z^{\mu-1} (1-z)^{\nu-1}\, du\, dz$$

$$= \left(\frac{e^{-u} \cdot u^{\mu + \nu - 1}}{\Gamma(\mu + \nu)} du \right) \left(\frac{1}{B(\mu, \nu)} z^{\mu - 1} (1 - z)^{\nu - 1} dz \right)$$

$$= [g_1 (u) du] [g_2 (z) dz], \quad \text{(say)},$$

where $g_1 (u) = \frac{1}{\Gamma(\mu + \nu)} e^{-u} u^{\mu + \nu - 1}, \ 0 < u < \infty$

and $g_2 (z) = \frac{1}{B(\mu, \nu)} z^{\mu - 1} (1 - z)^{\nu - 1}, \ 0 < z < 1$

From (*) and (**) we conclude that U and Z are independetnly distrib
U as a $\gamma(\mu + \nu)$ variate and Z as a $\beta_1 (\mu,\nu)$ variate.

Example 8·33. *If X and Y are independent Gamma variates parameters* μ *and* ν *respectively, show that*

$$U = X + Y, Z = \frac{X}{Y}$$

are independent and that U is a $\gamma(\mu + \nu)$ *variate and Z is a* $\beta_2 (\mu, \nu)$ *variat*

[Rajasthan Univ. B.Sc. (Hons.), 1

Solution. As in example 8·32, we have

$$dF(x, y) = \frac{1}{\Gamma(\mu) \Gamma(\nu)} e^{-(x+y)} x^{\mu - 1} y^{\nu - 1} dx\, dy, 0 < (x, y) < \infty$$

Since $u = x + y$ and $z = \frac{x}{y}$,

$$1 + z = 1 + \frac{x}{y} = \frac{u}{y} \Rightarrow y = \frac{u}{1 + z} \text{ and } x = \frac{uz}{1 + z} = u \left(1 - \frac{1}{1 + z} \right)$$

$$J = \frac{\partial (x, y)}{\partial (u, z)} = \frac{-u}{(1 + z)^2}$$

As x and y range from 0 to ∞, both u and z range from 0 to ∞. H
the joint probability differential of random variables U and Z becomes

$$dG(u, v) = \frac{1}{\Gamma(\mu) \Gamma(\nu)} e^{-u} \left(\frac{uz}{1 + z} \right)^{\mu - 1} \left(\frac{u}{1 + z} \right)^{\nu - 1} | J | dudz$$

$$= \left(\frac{e^{-u} u^{\mu + \nu - 1}}{\Gamma(\mu + \nu)} du \right) \left(\frac{1}{B(\mu, \nu)} \cdot \frac{z^{\mu - 1}}{(1 + z)^{\mu + \nu}} dv \right);$$

$$0 < u < \infty, 0 <$$

Hence U and Z are independently distributed, U as a $\gamma(\mu + \nu)$ va
and Z as a $\beta_2 (\mu, \nu)$ variate.

Remark. The above two examples lead to the following important re

If X is *a* $\gamma(\mu)$ variate and Y is an independent $\gamma(\nu)$ variate, then

(*i*) $X + Y$ is a $\gamma(\mu + \nu)$ variate, *i.e.*, the sum of two independent Ga
variates is also a Gamma variate.

(*ii*) $\dfrac{X}{Y}$ is a $\beta_2(\mu, \nu)$ variate, *i.e.*, the ratio of two independent Gamma variates is also a gamma variate.

(*iii*) $\dfrac{X}{X+Y}$ is a $\beta_1(\mu, \nu)$ variate.

Example 8·34. *Let X and Y have joint p.d.f.*

$$g(x, y) = \frac{e^{-(x+y)}\, x^3 y^4}{\Gamma 4\, \Gamma 5}, x > 0, y > 0$$
$$= 0, \textit{ elsewhere.}$$

Find (i) p.d.f. of $U = \dfrac{X}{X+Y}$, (*ii*) $E(U)$ and (*iii*) $E[U - E(U)]^2$

(Allahabad Univ. B.Sc. 1992)

Solution. Let $u = \dfrac{x}{x+y}$ and $v = x + y$

$\Rightarrow$ $x = uv, y = v - x = v - uv = v(1-u)$

Jacobian of transformation is

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

Hence joint p.d.f. of U and V becomes :

$$p(u, v) = g(x, y) \,.\, |J|$$
$$= \frac{1}{\Gamma 4\, \Gamma 5} e^{-v} \,.\, (uv)^3 [v(1-u)]^4 \times v$$
$$= \frac{1}{\Gamma 4\, \Gamma 5} e^{-v} \,.\, v^8 \,.\, u^3 (1-u)^4 \,;\, 0 \le u \le 1 \,;\, v > 0$$

$\left[\because\ u = \dfrac{x}{x+y} < 1 \text{ and since } x > 0, y > 0\right.$

we have $0 < u < 1$ and $v = x + y \ge 0$]

$$= \left[\frac{1}{\Gamma 9} e^{-v} \,.\, v^8\right] \left[\frac{\Gamma 9}{\Gamma 4\, \Gamma 5} u^3 (1-u)^4\right]; \quad 0 < u < 1, v > 0$$

$\Rightarrow$ $p(u, v) = p_1(v) \,.\, p_2(u)$ ...(*)

where $p_1(v) = \dfrac{1}{\Gamma 9} e^{-v} v^8 \,;\, v > 0$

$p_2(u) = \dfrac{\Gamma 9}{\Gamma 4\, \Gamma 5} u^3 (1-u)^4 \,;\, 0 < u < 1.$ ...(**)

From (*), we conclude that U and V are independently distributed and from (**), we conclude that

$$U = \frac{X}{X+Y} \sim \beta_1(4, 5).$$

i.e., U is a Beta variate of first kind with parameters (4, 5).

Aliter. We have

$$g(x, y) = \frac{1}{\Gamma 4 \,\Gamma 5} e^{-(x+y)} x^3 y^4$$

$$= \left[\frac{1}{\Gamma 4} e^{-x} x^3 \right] \left[\frac{1}{\Gamma 5} e^{-y} y^4 \right]$$

$$= g_1(x)\, g_2(y) \,;\, x > 0,\, y > 0$$

$\Rightarrow$ X and Y are independently distributed and $X \sim \gamma(4)$ and $Y \sim \gamma(5)$

Hence

$$U = \frac{X}{X+Y} \sim \beta_1(4, 5)$$

Now $$E(U) = \int_0^1 u \,.\, p_2(u)\, du = \frac{1}{B(4, 5)} \,.\, \int_0^1 u^4 (1-u)^4\, du$$

$$= \frac{1}{B(4, 5)} \,.\, B(5, 5)$$ [Using Beta integral]

$$= \frac{\Gamma 9}{\Gamma 4\, \Gamma 5} \times \frac{\Gamma 5\, \Gamma 5}{\Gamma 10} = \frac{\Gamma 9 \,.\, 4\,\Gamma 4}{\Gamma 4 \,.\, 9\,\Gamma 9} = \frac{4}{9}$$

$$E(U^2) = \frac{1}{B(4, 5)} \int_0^1 u^2 \,.\, u^3 (1-u)^4\, du$$

$$= \frac{1}{B(4, 5)} \times B(6, 5) = \frac{\Gamma 9}{\Gamma 4\, \Gamma 5} \times \frac{\Gamma 6\, \Gamma 5}{\Gamma 11}$$

$$= \frac{5 \times 4}{10 \times 9} = \frac{2}{9}$$

$$E[\,U - E(U)\,]^2 = E(U^2) - [\,E(U)\,]^2 = \frac{2}{9} - \frac{16}{81} = \frac{2}{81}$$

Example 8·35. *A random sample of size n is taken from a population with distribution:*

$$dP(x) = \frac{1}{\Gamma(\lambda)} e^{-x/a} \left(\frac{x}{a} \right)^{\lambda - 1} \frac{dx}{a} \,;\, 0 < x < \infty,\, a > 0,\, \lambda > 0$$

Find the distribution of the mean $\bar{X}$.

Solution. **[Delhi Univ. M.Sc. (OR), 1990]**

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} f(x)\, dx$$

$$= \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{tx} e^{-x/a} \left(\frac{x}{a} \right)^{\lambda - 1} \frac{dx}{a}$$

$$= \frac{1}{\Gamma(\lambda)\, a^{\lambda}} \int_0^{\infty} e^{-\left(\frac{1}{a}-t\right)x}\, x^{\lambda-1}\, dx$$

$$= \frac{1}{\Gamma(\lambda)\, a^{\lambda}} \cdot \frac{\Gamma(\lambda)}{\left(\frac{1}{a}-t\right)^{\lambda}} = (1-at)^{-\lambda} \qquad \ldots(*)$$

$\therefore \quad M_{\bar{X}}(t) = M_{(X_1+X_2 \ldots + X_n)/n}(t) = M_{X_1+X_2+\ldots+X_n}(t/n)$

$$[\because M_{cX}(t) = M_X(ct)]$$

$= M_{X_1}(t/n)\, M_{X_2}(t/n) \ldots M_{X_n}(t/n),$

(since $X_1, X_2, \ldots, X_n$ are independent).

Hence on using (*), we get

$$M_{\bar{X}}(t) = \left[\left(1-\frac{at}{n}\right)^{-\lambda}\right]^{n} = \left(1-\frac{t}{n/a}\right)^{-n\lambda}$$

which is the m.g.f. of a Gamma distribution (c.f. Remark 3, § 8·3·2). Hence by uniqueness theorem of mg.f., $\bar{X} \sim \gamma\,(n/a, n\,\lambda)$ with p.d.f.

$$g(\bar{x}) = \frac{(n/a)^{\lambda n}}{\Gamma(\lambda n)}\, e^{-n\bar{x}/a}\, (\bar{x})^{n\lambda-1}, 0 < \bar{x} < \infty$$

Example 8·36. *A sample of n values is drawn from a population whose probability density is ae^{-ax}, $(x \geq 0, a > 0)$. If $\bar{X}$ is mean of the sample, show that $na\bar{X}$ is a $\gamma(n)$ variate and prove that*

$$E(\bar{X}) = \frac{1}{a} \text{ and } S.E. \text{ of } \bar{X} = \frac{1}{a\sqrt{n}}$$

(Marathwada Univ. M.A., 1991)

Solution.

$$f(x) = ae^{-ax};\ 0 \leq x \leq \infty, a > 0$$

$$\therefore \quad M_X(t) = \int_0^{\infty} e^{tx} f(x)\, dx = a\int_0^{\infty} e^{-(a-t)x}\, dx$$

$$= a\left|\frac{e^{-(a-t)x}}{-(a-t)}\right|_0^{\infty} = \frac{a}{a-t}, \quad (a > t)$$

$$\because \quad an\bar{X} = an\left(\frac{X_1+X_2+\ldots+X_n}{n}\right) = a\,(X_1+X_2+X_n)$$

$$\therefore \quad M_{an\bar{X}}(t) = M_{a(X_1+X_2+\ldots+X_n)}(t) = M_{X_1+X_2+\ldots+X_n}(at)$$

$$= M_{X_1}(at)\,.\, M_{X_2}(at) \ldots M_{X_n}(at),$$

(since the sample values are independent).

$$\therefore M_{an\bar{X}}(t) = \prod_{i=1}^{n} M_{X_i}(at) = \left[M_{X_i}(at) \right]^n$$

(since $X_1, X_2, ..., X_n$ are identically distributed).

$$\therefore M_{an\bar{X}}(t) = \left(\frac{1}{1-t} \right)^n = (1-t)^{-n},$$ which is the m.g.f. of a $\gamma(n)$ variate.

Hence by uniqueness theorem of m.g.f., $an\bar{X}$ is a $\gamma(n)$ variate.

Since the mean and variance of a $\gamma(n)$ variate are equal, each being equal to n, we have

$$E(an\bar{X}) = n \Rightarrow an\,E(\bar{X}) = n, \text{ i.e., } E(\bar{X}) = \frac{1}{a}$$

and $$V(an\bar{X}) = n \Rightarrow a^2 n^2 V(\bar{X}) = n \text{ i.e., } V(\bar{X}) = \frac{1}{na^2}$$

Hence standard error (S.E.) of $\bar{X} = \sqrt{V(\bar{X})} = \dfrac{1}{a\sqrt{n}}$

Example 8·37. *Let $X \sim \beta_1(\mu, \nu)$ and $Y \sim \gamma(\lambda, \mu + \nu)$ be independent random variables, $(\mu, \nu, \lambda > 0)$. Find a p.d.f. for XY and identify its distribution.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

Solution. Since X and Y are independently distributed, their joint p.d.f. is given by

$$f(x, y) = \frac{1}{B(\mu, \nu)} \cdot x^{\mu-1}(1-x)^{\nu-1} \times \frac{\lambda^{\mu+\nu}}{\Gamma(\mu+\nu)} e^{-\lambda y} y^{\mu+\nu-1};$$

$$0 < x < 1, 0 < y < \infty$$

Let us transform to the new variables U and Z by the transformation

$xy = u, x = z$ i.e, $x = z$ and $y = u/z$

Jacobian of transformation J is given by

$$J = \frac{\partial(x, y)}{\partial(u, z)} = \begin{vmatrix} 0 & 1 \\ \frac{1}{z} & \frac{-u}{z^2} \end{vmatrix} = -\frac{1}{z}$$

Thus the joint p.d.f. of U and Z becomes

$$g(u, z) = \frac{\lambda^{\mu+\nu}}{B(\mu, \nu)\Gamma(\mu+\nu)} \cdot (z)^{\mu-1}(1-z)^{\nu-1} e^{-\lambda u/z} \left(\frac{u}{z} \right)^{\mu+\nu-1} |J|;$$

$$0 < u < \infty, 0 < z < 1$$

Integrating w.r.t. z in the range $0 < z < 1$, the marginal p.d.f. of U is given by

$$= \frac{\lambda^{\mu+\nu} u^{\mu+\nu-1}}{\Gamma(\mu)\Gamma(\nu)} \int_0^1 \frac{(1-z)^{\nu-1} e^{-\lambda u/z}}{z^{\nu+1}} dz$$

$$= \frac{\lambda^{\mu+\nu} u^{\mu+\nu-1}}{\Gamma(\mu)\Gamma(\nu)} \int_0^1 \frac{1}{z^2} \left(\frac{1}{z} - 1 \right)^{\nu-1} e^{-\lambda u/z} dz$$

Thus

$$g(u) = \frac{\lambda^{\mu+\nu} u^{\mu+\nu-1}}{\Gamma(\mu)\Gamma(\nu)} \int_{\infty}^{0} t^{\nu-1} e^{-\lambda u(1+t)} (-dt) \qquad \left[\frac{1}{z} - 1 = t\right]$$

$$= \frac{\lambda^{\mu+\nu} u^{\mu+\nu-1} e^{-\lambda u}}{\Gamma(\mu)\Gamma(\nu)} \int_0^{\infty} e^{-\lambda ut} t^{\nu-1} dt$$

$$= \frac{\lambda^{\mu+\nu} u^{\mu+\nu-1} e^{-\lambda u}}{\Gamma(\mu)\Gamma(\nu)} \frac{\Gamma(\nu)}{(\lambda u)^{\nu}}$$

$$= \frac{\lambda^{\mu}}{\Gamma(\mu)} \cdot e^{-\lambda u} u^{\mu-1}, \; 0 < u < \infty$$

Hence $U = XY$ is distributed as a gamma variate with parameters λ and μ, *i.e.*, $XY \sim \gamma(\lambda, \mu)$.

Example 8·38. Let $p \sim \beta_1(a, b)$ *where a and b are positive integers. After one observes p, one secures a coin for which the probability of heads is p. This coin is flipped n times. Let X denote the number of heads which result. Find* $P(X = k)$ *for* $k = 0, 1, 2, \ldots, n$. *Express the answer in terms of binomial co-efficients.*

Solution. Since $p \sim \beta_1(a, b)$, its p.d.f. is given by

$$P(p) = \frac{1}{B(a, b)} \, p^{a-1} (1-p)^{b-1}, \; 0 < p < 1$$

$P(X = k \mid$ the probability of success in a single trial is $p)$

$$= \binom{n}{k} p^k q^{n-k}, \; q = 1 - p$$

$$P(X = k) = \int_0^1 P(p) P(X = k \mid p) \, dp$$

$$= \int_0^1 \frac{1}{B(a, b)} \cdot p^{a-1} (1-p)^{b-1} \cdot \binom{n}{k} p^k (1-p)^{n-k} dp$$

$$= \frac{\binom{n}{k}}{B(a, b)} \int_0^1 p^{a+k-1} (1-p)^{n+b-k-1} dp$$

$$= \frac{\binom{n}{k} B(a+k, n+b-k)}{B(a, b)} \qquad \ldots(1)$$

We have

$$\frac{1}{B(m, n)} = \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} = \frac{(m+n-1)!}{(m-1)!\,(n-1)!} = \frac{mn}{m+n} \binom{m+n}{m} \qquad \ldots(2)$$

$$\therefore \qquad P(X=k) = \frac{\binom{n}{k}\dfrac{ab}{a+b}\binom{a+b}{a}}{\dfrac{(a+k)(n+b-k)}{(n+a+b)}\binom{n+a+b}{a+k}}$$

$$= \frac{\binom{n}{k}\binom{a+b}{a}}{\binom{n+a+b}{a+k}} \cdot \frac{ab\,(n+a+b)}{(a+b)(a+k)(n+b-k)}$$

Example 8·39. *Given the Incomplete Beta Function,*

$$B_x(l, m) = \int_0^x x^{l-1}(1-x)^{m-1}\,dx$$

and $\quad I_x(l, m) = B_x(l, m)/B(l, m)$, *show that*

$\quad I_x(l, m) = 1 - I_{1-x}(m, l)$.

Solution. We have

$$I_x(l, m)\,B(l, m) = B_x(l, m) = \int_0^x x^{l-1}(1-x)^{m-1}\,dx \qquad \ldots(*)$$

$$= \int_0^1 x^{l-1}(1-x)^{m-1}\,dx - \int_x^1 x^{l-1}(1-x)^{m-1}\,dx$$

$$= B(l, m) - \int_x^1 x^{l-1}(1-x)^{m-1}\,dx$$

In the integral, put $1 - x = y$, then

$$I_x(l, m)\,.\,B(l, m) = B(l, m) - \int_{1-x}^0 (1-y)^{l-1}\,y^{m-1}\,(-dy)$$

$$= B(l, m) - \int_0^{1-x} y^{m-1}(1-y)^{l-1}\,dy$$

$$= B(l, m) - B_{1-x}(m, l) = B(l, m) - I_{1-x}(m, l)\,B(m, l)$$

[From (*)]

Since $\quad B(l, m) = B(m, l)$, we get on dividing throughout by $B(l, m)$

$\quad I_x(l, m) = 1 - I_{1-x}(m, l)$

EXERCISE 8(d)

1. (*a*) Suppose the frequency function of a random variable is given by

$$f(x) = \begin{cases} \dfrac{x^k e^{-x}}{k!}, & \text{for } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

where k is non-negative integer.

(*i*) Find the moment generating function of this distribution.

(*ii*) Determine the mean and variance of this distribution. using moment generating function.

(*b*) If X is a Gamma variate with parameter λ, obtain its m.g.f. Hence deduce that the m.g.f. of standard gamma variate tends to $e^{t^2/2}$ as $\lambda \to \infty$. Also interpret the result. **[Delhi Univ. B.A. (Stat. Hons.), 1988, '82]**

(*c*) X and Y are two independent gamma variates, with parameters l and m. Prove that $(X + Y)$ is a gamma variate with parameter $(l + m)$.

(*d*) If $X_1, X_2, ..., X_n$ are independent and identically distributed gamma random variables, what is the distribution of $X_1 + X_2 + ... + X_n$?

[Delhi Univ. B. Sc. (Maths. Hons.), 1988]

(*e*) Cosider a random variable X with the following *p.d.f.*

$$f(x) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} x^{\alpha - 1} e^{-x/\beta} \; ; 0 < x < \infty; \alpha, \beta > 0$$

Find the moment generating function of X.

Let the random variable X with above p.d.f. be defined as $X \sim Ga\ (\alpha, \beta)$. Then prove the following theorems :

Theorem 1 : If X and Y are independent and $X \sim Ga\ (\alpha_1, \beta)$ and $Y \sim Ga\ (\alpha_2, \beta)$, then $X + Y \sim Ga\ (\alpha_1 + a_2, \beta)$

Theorem 2 : If X and Y are independent and $X \sim Ga\ (\alpha_1, \beta)$ and $X + Y \sim Ga\ (\alpha_1 + \alpha_2, \beta)$, then $Y \sim Ga\ (\alpha_2, \beta)$ **(Mysore Univ. B.Sc. 1993)**

2. (*a*) Define the Beta variate of first kind. Obtain its mean and variance. Also define the Beta variate of second kind and state its relation with Gamma variates. **(Nagpur Univ. B.Sc. 1993)**

(*b*) Write down the Beta probability functions of the first kind and the second kind with parameters μ and ν. Show that a Beta variate of the first kind can be obtained by a transformation of a Beta variate of the second kind.

(*c*) If X has a Beta distribution, can $E\ (1/X)$ be unity ?

Ans. $X \sim B\ (m, n)$; $E\ (1/X) = \dfrac{m + n - 1}{n - 1} > 1$. No.

3. Let $X \sim \gamma\ (\lambda, a)$ and $Y - \gamma\ (\lambda, b)$, be independent random variables. Show that :

$$E\left[\frac{X}{X + Y}\right] = \frac{E(X)}{E\ (X + Y)}$$

Hint. Since $X \sim \gamma\ (\lambda, a)$ and $Y \sim \gamma\ (\lambda, b)$ are independent, $U = X/\ (X + Y)$ and $V = (X + Y)$ are independent. Hence

$$E\ (UV) = E\ (U)\ E\ (V) \;\Rightarrow\; E\ (X) = E\left[\frac{X}{X + Y}\right] \cdot E\ (X + Y).$$

4. (*a*) If $X \sim \gamma(\lambda, \mu)$ and $\gamma - \gamma(\lambda, \nu)$ are independent random variables, show that

$$\frac{X}{Y} \sim \beta_2(\mu, \nu)$$

(*b*) X and Y are independent Gamma variates find the distribution of $X/(X+Y)$.

(*c*) If X is random variable having as its rth moment

$$\mu'_r = \frac{(k+r)!}{k!}$$

k being a positive integer, show that its probability density function is

$$f(x) = \begin{cases} \dfrac{x^k}{k!} e^{-x}, & x > 0 \\ 0 & x < 0. \end{cases}$$

(*d*) If the r.v. X is such that

$$E(X^n) = (n+k)!\ k^n / k!; \ n = 1, 2, 3, \ldots$$

k being a positive integer, find the p.d.f. of X.

Ans. $X \sim \gamma\left(\frac{1}{k}, k+1\right)$ **[Delhi Univ. M.A. (Econ.), 1987]**

5. If X and Y are independent Gamma variates with parameters μ and ν respectively, show that the variables

$$U = X + Y \text{ and } V = \frac{X - Y}{X + Y}$$

are independent variables.

6. If X and Y are independent Gamma variates with parameter λ and μ respectively, show that the variables:

(*a*) $$U = X + Y \text{ and } V = \frac{X}{X+Y}$$

are independently distributed and identify their distributions.

[Delhi Univ. B.Sc. (Stat Hons.) 1991]

(*b*) $U = X + Y$ and $V = X/Y$ are independently distributed, $U \sim \gamma(\lambda + \mu)$ and $V \sim \beta_2(\lambda, \mu)$.

7. A simple sample of n values $x_1, x_2, \ldots, x_n$ is drawn from the population:

$$dP(x) = \frac{1}{\Gamma(n)} e^{-x} x^{n-1}\, dx, 0 \le x < \infty$$

If $\bar{x}$ is the mean of the sample, find the distribution of $n\bar{x}$. Hence find the mean and variance of the distribution.

8. (*a*) show that for a $\gamma(\lambda)$ distribution,

$$\frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{1}{\sqrt{\lambda}} = \frac{1}{2}\frac{\mu_3}{\sigma^3}.$$

Show that the excess of kurtosis of the distribution is $6/\lambda$.

(*b*) Show that the mean value of the positive square root of a $\gamma(\lambda, n)$ variate is $\Gamma(n+\frac{1}{2}) / [\sqrt{\lambda}\, \Gamma(n)]$.

Hence prove that the mean deviation of a $N(\mu, \sigma^2)$ variate from its mean is $\sigma\sqrt{2/\pi}$

[Gauhati Univ. M.A. (Eco)., 1991; Delhi Univ. B.Sc. (Stat Hons.), 1989]

Hint. Proceed as in Example 8·31.

9. Show that the mean value of the positive square root of $\beta(\mu, \nu)$ variate is

$$\frac{\Gamma\left(\mu+\frac{1}{2}\right)\Gamma(\mu+\nu)}{\Gamma(\mu)\,\Gamma\left(\mu+\nu+\frac{1}{2}\right)}$$

10. (*a*) For the distribution :

$$dP(x) = \frac{1}{B(\mu, \nu)} \frac{x^{\mu-1}}{(1+x)^{\mu+\nu}}\,;\ 0 < x < \infty, \nu > 2$$

show that variance is $\dfrac{\mu(\mu+\nu-1)}{(\nu-1)^2(\nu-2)}$.

Find also the mode and μ_r' (about origin). Also show that harmonic mean is $(\mu-1)/\nu$.

(*b*) Find the arithmetic mean, harmonic mean and variance of a Beta distribution of first kind with parameter μ and ν. Verify that A.M. > H.M.

Also prove that if G is the geometric mean, then

$$\log G = \frac{1}{B(\mu, \nu)} \frac{\partial}{\partial \nu} \beta(\mu, \nu) = \frac{\partial}{\partial \nu}\left[\log\sqrt{2} - \log\Gamma(\mu+\nu)\right]$$

11. Given the Beta distribution in the following form :

$$p(x) = \frac{1}{B(\alpha+1, \lambda+1)} \cdot x^{\alpha}(1-x)^{\lambda}\,;\ \alpha > -1, \lambda > -1, 0 \le x \le 1$$

find its variance.

Also find the distribution of (*i*) $\dfrac{1}{X}$, (*ii*) $\dfrac{1-X}{X}$.

12. If X is a normal variate with mean μ and standard deviation σ, find the mean and variance of Y defined by

$$Y = \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$$ **(Meerut Univ. B. Sc., 1993)**

8·6. The Exponential Distribution. A continuous random variable X assuming non-negative values is said to have an exponential distribution with parameter $\theta > 0$, if its p.d.f. is given by

$$f(x) = \begin{cases} \theta \cdot e^{-\theta x}, & x \ge 0 \\ 0, & \text{otherwise} \end{cases} \qquad \ldots(8\cdot24)$$

The cumulative distribution function $F(x)$ is given by

$$F(x) = \int_0^x f(u)\, du = \theta \int_0^x \exp(-\theta u)\, du$$

$$F(x) = \begin{cases} 1 - \exp(-\theta x), & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \ldots [8\cdot24(a)]$$

8·6·1. Moment Generating Function of Exponential Distribution

$$M_X(t) = E(e^{tX}) = \theta \int_0^\infty e^{tx}\, e^{-\theta x}\, dx$$

$$= \theta \int_0^\infty \exp\{-(\theta - t)x\}\, dx = \frac{\theta}{(\theta - t)},\ \theta > t$$

$$= \left(1 - \frac{t}{\theta}\right)^{-1} = \sum_{r=0}^{\infty} \left(\frac{t}{\theta}\right)^r$$

$\therefore$ $$\mu_r' = E(X^r) = \text{Coefficient of } \frac{t^r}{r!} \text{ in } M_X(t)$$

$$= \frac{r!}{\theta^r};\ r = 1, 2, \ldots$$

$\therefore$ $$Mean = \mu_1' = \frac{1}{\theta}$$

$$Variance = \mu_2 = \mu_2' - \mu_1'^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$$

Theorem. If $X_1, X_2, \ldots, X_n$ are independent random variables, X_i having an exponential distribution with parameter θ_i; $i = 1, 2, \ldots, n$; then $Z = \min(X_1, X_2, \ldots, X_n)$ has exponential distribution with parameter $\sum_{i=1}^{n} \theta_i$.

[Delhi Univ. B.Sc. (Stat. Hons.), 1986]

Proof. $$G_Z(z) = P(Z \leq z) = 1 - P(Z > z)$$

$$= 1 - P[\min(X_1, X_2, \ldots, X_n) > z]$$

$$= 1 - P[X_i > z;\ i = 1, 2, \ldots, n]$$

$$= 1 - \prod_{i=1}^{n} P(X_i > z) = 1 - \prod_{i=1}^{n} [1 - P(X_i \leq z)]$$

$$= 1 - \prod_{i=1}^{n} [1 - F_{X_i}(z)]$$

where F is the distribution function of X_i.

$$= 1 - \prod_{i=1}^{n} \left[1 - \left(1 - e^{-\theta_i z}\right)\right] \quad \text{[c.f. 8·24 (a)]}$$

$$= \begin{cases} 1 - \exp\left\{ (-\sum_{i=1}^{n} \theta_i) z \right\}, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\therefore \quad g_Z(z) = \begin{cases} (\sum_{i=1}^{n} \theta_i) \exp\left\{ (-\sum_{i=1}^{n} \theta_i) z \right\}, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$

$\Rightarrow Z = \min(X_1, X_2, ..., X_n)$ is an exponential variate with **parameter** $\sum_{i=1}^{n} \theta_i$.

Cor. If X_i ; $i = 1, 2, ..., n$ are identically distributed, following exponential distribution with parameter θ, then $Z = \min(X_1, X_2, ..., X_n)$ is also exponentially distributed with parameter $n\,\theta$.

Example 8·40. *Show that the exponential distribution "lacks memory", i.e., if X has an exponential distribution, then for every constant $a \geq 0$, one has $P(Y \leq x \mid X \geq a) = P(X \leq x)$ for all x, where $Y = X - a$.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1989; Calicut Univ. B.Sc. (Main Stat.), 1991]

Solution. The p.d.f. of the exponential distribution with parameter θ is

$$f(x) = \theta\, e^{-\theta x} ; \theta > 0, 0 < x < \infty$$

We have

$$\begin{aligned} P(Y \leq x \cap X \geq a) &= P(X - a \leq x \cap X \geq a) \qquad (\because Y = X - a) \\ &= P(X \leq a + x \cap X \geq a) = P(a \leq X \leq a + x) \\ &= \theta \int_a^{a+x} e^{-\theta x}\, dx = e^{-a\theta}(1 - e^{-\theta x}) \end{aligned}$$

and

$$P(X \geq a) = \theta \int_a^{\infty} e^{-\theta x}\, dx = e^{-a\theta}$$

$$\therefore \quad P(Y \leq x \mid X \geq a) = \frac{P(Y \leq x \cap X \geq a)}{P(X \geq a)} = 1 - e^{-\theta x} \qquad ...(*)$$

Also

$$P(X \leq x) = \theta \int_0^x e^{-\theta x} dx = 1 - e^{-\theta x} \qquad ...(**)$$

From (*) and (**), we get

$$P(Y \leq x \mid X \geq a) = P(X \leq x)$$

i.e., exponential distribution lacks memory.

Example 8·41. *X and Y are independent with a common p.d.f. (exponential):*

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Find a p.d.f. for $X - Y$. **[Delhi Univ. B.Sc. (Stat. Hons.), 1988, '85]**

Solution. Since. X and Y are independent and identically distributed (*i.i.d.*), their joint *p.d.f.* is given by

$$fxy\,(x, y) = \begin{cases} e^{-(x+y)}; \; x > 0, \; y > 0 \\ 0 \end{cases}$$

otherwise

$$\text{Let} \quad \begin{cases} u = x - y \\ v = y \end{cases} \Rightarrow \begin{cases} x = u + v \\ y = v \end{cases} \qquad \ldots(1)$$

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1$$

Thus the joint p.d.f. of U and V becomes

$$g(u, v) = e^{-(u+2v)}; \; v > 0, \; -\infty < u < \infty$$

(1) $\Rightarrow$ $\quad u = x - v \Rightarrow v = x - u$

Thus $\quad v > -u$ if $-\infty < u < 0$

and $\quad v > 0$ if $u > 0$

For $\quad -\infty < u < 0$,

$$g(u) = \int_{-u}^{\infty} g(u, v)\, dv = \int_{-u}^{\infty} e^{-(u+2v)}\, dv = e^{-u} \left| \frac{e^{-2v}}{-2} \right|_{-u}^{\infty} = \frac{1}{2} e^{u}$$

and for u > 0,

$$g(u) = \int^{\infty} g(u, v)\, dv = e^{-u} \left| \frac{e^{-v}}{-2} \right|_{-u}^{\infty} = \frac{1}{2} e^{-u^2}$$

Hence the p.d.f of $U = X - Y$ is given by

$$g(u) = \begin{cases} \dfrac{1}{2} e^{u}, \; -\infty < u < 0 \\ \dfrac{1}{2} e^{-u}, \; u > 0 \end{cases}$$

These results can be combined to give

$$g(u) = \frac{1}{2} e^{-|u|}, \; -\infty < u < \infty$$

which is the p.d.f. of standard Laplace distribution (*c.f.* § 8·7).

Aliter.

$$M_X(t) = \int_0^{\infty} e^{tx} f(x)\, dx = \int_0^{\infty} e^{-(1-t)x}\, dx = \left| \frac{e^{-(1-t)x}}{-(1-t)} \right|_0^{\infty} = \frac{1}{1-t}, \; t < 1$$

$\therefore$ Characteristic function of X is

$$\varphi_X(t) = \frac{1}{1 - it} = \varphi_Y(t),$$

(since X and Y are identically distributed.)

$\therefore \; \varphi_{X-Y}(t) = \varphi_{X+(-Y)}(t) = \varphi_{X}(t)\, \varphi_{-Y}(t) \qquad (\because X, Y$ are independent)

$$= \varphi_X(t) \cdot \varphi_Y(-t) = \frac{1}{(1-it)(1+it)} = \frac{1}{1+t^2}$$

which is the characteristic function of the Laplace distribution, (c.f. § 8·7)

$$g(u) = \frac{1}{2} e^{-|u|}, -\infty < u < \infty \qquad ...(*)$$

Hence by the uniqueness theorem of characteristic functions, $U = X - Y$ has the p.d.f. given in (*).

8·7. Laplace (Double Exponential) Distribution. A continuous random variable X is said to follow standard Laplace distribution if its p.d.f. is given by

$$f(x) = \frac{1}{2} e^{-|x|}, -\infty < x < \infty \qquad ...(8·25)$$

Characteristic function is given by

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx = \frac{1}{2} \int_{-\infty}^{\infty} e^{itx} . e^{-|x|}\, dx$$

$$= \frac{1}{2}\left[\int_{-\infty}^{\infty} \cos tx . e^{-|x|}\, dx + i \int_{-\infty}^{\infty} \sin tx . e^{-|x|}\, dx\right.$$

$$= \frac{1}{2} . 2 \int_{0}^{\infty} \cos tx . e^{-|x|}\, dx,$$

Since the integrands in the first and second integrals are even and odd function of x respectively.

$$\therefore \qquad \varphi_X(t) = \int_0^{\infty} e^{-x} \cos tx\, dx$$

$$= 1 - t^2 \int_0^{\infty} e^{-x} \cos tx\, dx \quad \text{(on integration by parts)}$$

$$= 1 - t^2\, \varphi_X(t)$$

$$\Rightarrow \qquad \varphi_X(t) = \frac{1}{1+t^2} \qquad ...(8·25\ a)$$

The mean of this distribution is zero, standard deviation is $\sqrt{2}$ and mean deviation about mean is 1.

Remark. Generalised Laplace Distribution. A continuous r.v. X is said to have Laplace distribution with two parameters λ and μ if its p.d.f. is given by

$$f(x) = \frac{1}{2\lambda} \exp\left[- |x - \mu|/\lambda\right], -\infty < x < \infty\, ; \lambda > 0 \qquad ...(8·26)$$

Taking $U = \dfrac{X - \mu}{\lambda}$, in (8·26) we obtain the p.d.f. of standard Laplace variate given in (8·25).

Moments. The rth moment about origin is given by

$$\mu'_r = E(X^r) = \frac{1}{2\lambda}\int_{-\infty}^{\infty} x^r \exp\left(\frac{-|x-\mu|}{\lambda}\right) dx$$

$$= \frac{1}{2}\int_{-\infty}^{\infty} (z\lambda + \mu)^r \exp(-|z|)\, dz, \qquad \left[z = \frac{x-\mu}{\lambda}\right]$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\left[\sum_{k=0}^{r}\binom{r}{k}(z\lambda)^k \mu^{r-k}\right]\exp(-|z|)\, dz$$

$$= \frac{1}{2}\sum_{k=0}^{r}\left[\binom{r}{k}\lambda^k \mu^{r-k}\int_{-\infty}^{\infty} z^k \exp(-|z|)\, dz\right]$$

$$= \frac{1}{2}\sum_{k=0}^{r}\left[\binom{r}{k}\lambda^k \mu^{r-k}\left\{\int_{-\infty}^{0} z^k e^{(-|z|)}\, dz + \int_{0}^{\infty} z^k e^{-|z|}\, dz\right\}\right]$$

$$= \frac{1}{2}\sum_{k=0}^{r}\left[\binom{r}{k}\lambda^k \mu^{r-k}\left\{(-1)^k\int_{0}^{\infty} e^{-z} z^k\, dz + \int_{0}^{\infty} e^{-z} z^k\, dz\right\}\right]$$

$$= \frac{1}{2}\sum_{k=0}^{r}\left[\binom{r}{k}\lambda^k \mu^{r-k}\,\Gamma(k+1)\,\{(-1)^k + 1\}\right]$$

$$\Rightarrow \qquad \mu'_r = \frac{1}{2}\sum_{k=0}^{r}\left[\binom{r}{k}\lambda^k \mu^{r-k}\, k!\,\{1+(-1)^k\}\right] \qquad \text{...(8·26}a\text{)}$$

$\therefore$ Mean $= \mu'_r = \mu_1' = \mu$ and $\mu_2' = \mu^2 + 2\lambda^2$

$\therefore$ $\sigma_X^2 = \mu_2' - \mu_1'^2 + 2\lambda^2$.

Similarly we can obtain higher order moments from (8·26 a) and hence the values of β_1 and β_2 can be obtained.

The characteristic function of (8·26) can be obtained exactly similarly as we obtained the characteristic function of standard Cauchy distribution, c.f.§ 8.9.

8·8. Weibul Distribution. A random variable X has a Weibul distribution with three parameters c (> 0), α(> 0) and μ if the r.v.

$$Y = \left(\frac{X-\mu}{\alpha}\right)^c \qquad \text{...(}i\text{)}$$

has the exponential distribution with p.d.f.

$$p_Y(y) = e^{-y},\ y > 0 \qquad \text{...(}ii\text{)}$$

The p.d.f. of X is given by

$$p_X(x) = c\,\alpha^{-1}\left(\frac{x-\mu}{\alpha}\right)^{c-1} \exp\left[-\left(\frac{x-\mu}{\alpha}\right)^c\right],\quad x > \mu,\ c > 0 \qquad ...(iii)$$

The standard Weibul distribution is obtained on taking $\alpha = 1$ and $\mu = 0$, so that the p.d.f. of *standard Weibul distribution* which depends only on a single parameter c is given by

$$p_X(x) = c\,x^{c-1} \,.\, \exp(-x^c)\,;\ x > 0,\ c > 0 \qquad ...(iv)$$

8·8·1. Moments of Standard Weibul Distribution (iv)

For standard Weibul distribution, $(\alpha = 1, \mu = 0)$, from (i), we get $Y = X^c$ which has the exponential distribution (ii). We have

$$\mu'_r = E(X^r) = E(Y^{1/c})^r = E(Y^{r/c})$$

$$= \int_0^{\infty} e^{-y} \,.\, y^{r/c}\, dy \qquad [\because\ Y \text{ has p.d.f. } (ii)]$$

$$\Rightarrow \qquad \mu'_r = \Gamma\left(\frac{r}{c} + 1\right) \qquad ...(v)$$

$$\therefore \qquad \text{Mean} = E(X) = \Gamma\left(\frac{1}{c} + 1\right),$$

and $\quad \text{Var}(X) = E(X^2) = [E(X)]^2$

$$= \Gamma\left(\frac{2}{c} + 1\right) - \left[\Gamma\left(\frac{1}{c} + 1\right)\right]^2$$

Similarly, we can obtain expressions for higher order moments and hence for β_1 and β_2. For large c, the mean is approximated by

$$E(X) \simeq 1 - \frac{\gamma}{c} + \frac{1}{2c^2}\left(\frac{\pi^2}{6} + \gamma^2\right)$$

$$= 1 - 0{\cdot}57722\ c^{-1} + 0{\cdot}98905\ c^{-2}$$

where $\gamma = 0{\cdot}57722$ is Euler's constant.

The distribution is named after Waloddi Weibul, a Swedish physicist, who used it in 1939 to represent the distribution of the breaking strength of materials. Kao, J.H.K. (1958-59) advocated the use of this distribution in reliability studies and quality control work. It is also used as a tolerance distribution in the analysis of quantum response data.

8·8·2. Characterisation of Weibul Distribution. Dubey, S.D. (1968) has obtained the following result :

"Let X_i $(i = 1, 2, ..., n)$ be i.i.d. random variables. Then min $(X_1, X_2, ..., X_n)$ has a Weibul distribution if and only if the common distribution of X_i's is a Weibul distribution."

Proof. Let X_i, $(i = 1, 2, ..., n)$ be *i.i.d. r.v.* each with Weibul distribution (iii) and let $Y = \min(X_1, X_2, ..., X_n)$. Then

$$P(Y > y) = P[\min(X_1, X_2, ..., X_n) > y]$$

$$= P\left[\bigcap_{i=1}^{n} X_i > y\right]$$

$$= \prod_{i=1}^{n} P(X_i > y) = \left[P(X_i > y)\right]^n \qquad \ldots(*)$$

since X_i's are *i.i.d. r.v.*'s.

$$\text{Now} \quad P(X_i > y) = \int_y^{\infty} c\,\alpha^{-1}\left(\frac{x-\mu}{\alpha}\right)^{c-1} . \exp\left[-\left(\frac{x-\mu}{\alpha}\right)^c\right] dx$$

$$= \int_{\left(\frac{y-\mu}{\alpha}\right)^c}^{\infty} e^{-t}\,dt \qquad \left[t = \left(\frac{x-\mu}{\alpha}\right)^c\right]$$

$$= \exp\left[-\left(\frac{y-\mu}{\alpha}\right)^c\right]$$

Substituting in (*), we get

$$P(Y > y) = \left[\exp\left\{-\left(\frac{y-\mu}{\alpha}\right)^c\right\}\right]^n$$

$$= \exp\left[-n\left(\frac{y-\mu}{\alpha}\right)^c\right]$$

$$= \exp\left[-\left\{\frac{n^{1/c}(y-\mu)}{\alpha}\right\}^c\right]$$

This implies that Y has the same Weibul distribution as X_i's with the difference that the parameter α is replaced by $\alpha\, n^{-1/c}$.

8·8·3. Logistic Distribution. A continuous *r.v.* X is said to have a Logistic distribution with parameters α and β, if its distribution function is of the form:

$$F_X(x) = [\,1 + \exp\{-(x-\alpha)/\beta\}\,]^{-1}, \quad \beta > 0 \qquad \ldots(8·26\,b)$$

$$= \frac{1}{2}\left[1 + \tanh\left\{\frac{1}{2}(x-\alpha)/\beta\right\}\right] \quad ; \ \beta > 0 \qquad \ldots(8·26\,c)$$

(See Remark 1 on page 8·94).

The p.d.f. of Logistic distribution with parameters α and β (> 0) is given by

$$f(x) = \frac{d}{dx}(F(x))$$

$$= \frac{1}{\beta}[\,1 + \exp\{-(x-\alpha)/\beta\}\,]^{-2}\,[\exp\{-(x-\alpha)/\beta\}\,] \qquad \ldots(8·26\,d)$$

$$= \frac{1}{4\beta}\,\text{sech}^2\left\{\frac{1}{2}(x-\alpha)/\beta\right\} \qquad \ldots(8·26\,e)$$

The p.d.f. of standard Logistic variate $Y = (X-\alpha)/\beta$, is given by:

$$g_Y(y) = f(x)\,.\left|\frac{dx}{dy}\right|$$
$$= e^{-y}\left(1+e^{-y}\right)^{-2}\ ; -\infty < y < \infty \qquad \text{...(8·26 } f)$$
$$= \frac{1}{4}\,\text{sech}^2\left(\frac{1}{2}y\right);\ -\infty < y < \infty \qquad \text{...(8·26 } g)$$

The distribution function of Y is :

$$G_Y(y) = (1+e^{-y})^{-1}\ ;\ -\infty < y < \infty \qquad \text{...(8·26 } h)$$

Logistic distribution is extensively used as growth function in population and demographic studies and in time series analysis. Theoretically, Logistic distribution can be obtained as :

(*i*) The limiting distribution (as $n \to \infty$) of the standardised mid range, (average of the smallest and the largest sample observations), in random samples of size n

(*ii*) A mixture of extreme value distributions.

Moment Generating Function. The m.g.f. of standard Logistic variate Y is given by:

$$\mu_Y(t) = E\left(e^{tY}\right) = \int_{-\infty}^{\infty} e^{ty}\,.\,g(y)\,dy$$
$$= \int_{-\infty}^{\infty} e^{ty}\,.\,e^{-y}(1+e^{-y})^{-2}\,dy$$
$$= \int_{-\infty}^{\infty} e^{ty}\,e^{-y}\left(\frac{1+e^y}{e^y}\right)^{-2} dy$$
$$= \int_{-\infty}^{\infty} e^{ty}\,.\,e^y(1+e^y)^{-2}\,dy$$

Put $\quad z = (1+e^y)^{-1} \Rightarrow e^y = \dfrac{1}{z} - 1 = \dfrac{1-z}{z}$

$$\therefore M_Y(t) = \int_1^0 \left(\frac{1-z}{z}\right)^t.(-dz) = \int_0^1 z^{-t}(1-z)^t\,dz$$
$$= \beta(1-t,\ 1+t),\quad 1-t > 0$$
$$= \Gamma(1-t)\,\Gamma(1+t)/\Gamma 2$$
$$= \Gamma(1-t)\ \Gamma(1+t)$$
$$= \pi t \text{ cosec } \pi t \quad ; t < 1 \qquad \text{...(8·26 } i)$$
$$= 1 + \frac{\pi^2 t^2}{6} + \frac{7}{360}\pi^4 t^4 + \qquad \text{...(*)}$$

(See Remark 2 below.)

$\therefore\ E(Y) = \text{Coefficient of } t \text{ in } (*) = 0$

$\Rightarrow \qquad \text{Mean} = 0$

$$\therefore \quad \mu_2 = E(Y^2) = \text{Coefficient of } \frac{t^2}{2!} \text{ in } (*) = \frac{\pi^2}{3}$$

$$\mu_3 = E(Y^3) = 0$$

$$\mu_4 = E(Y^4) = \text{Coefficient of } \frac{t^4}{4!} \text{ in } (*) = \frac{7}{15}\pi^4$$

Hence for standard Logistic distribution :

$$\text{Mean} = 0\,,\ \text{Variance} = \mu_2 = \pi^2/3,$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0, \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{7 \times 9}{15} = 4{\cdot}2$$

Remarks 1. We have :

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

$$\Rightarrow \quad 1 + \tanh x = \frac{2}{1 + e^{-2x}} \Rightarrow \frac{1}{2}[1 + \tanh x] = (1 + e^{-2x})^{-1}$$

2. $$x \operatorname{cosec} x = 1 + \frac{x^2}{6} + \frac{7}{360}x^4 + \dots$$

Proof. $$x \operatorname{cosec} x = \frac{x}{\sin x} = \frac{x}{\left[x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} = \dots\right]}$$

$$= \left[1 - \left(\frac{x^2}{6} - \frac{x^4}{120} + \frac{x^6}{7!} \dots\right)\right]^{-1}$$

$$= 1 + \left(\frac{x^2}{6} - \frac{x^4}{120} + \dots\right) + \left(\frac{x^2}{6} - \frac{x^4}{120} + \dots\right)^2 + \dots$$

$$= 1 + \frac{x^2}{6} + x^4\left(\frac{1}{36} - \frac{1}{120}\right) + \dots$$

$$= 1 + \frac{x^2}{6} + \frac{7}{360}x^4 + \dots$$

3. We have:

$$g(y) = e^{-y}\left(1 + \frac{1}{e^y}\right)^{-2} = e^y\left(1 + e^y\right)^{-2} = g(-y)$$

$\Rightarrow$ The probability curve of Y is symmetric about the line $y = 0$.

Since p.d.f. $g(y)$ is symmetric about origin $(y = 0)$, all odd order moments about origin are zero *i.e.*,

$$\mu'_{2r+1} = E\left(Y^{2r+1}\right) = 0, \quad r = 0, 1, 2, \dots$$

In particular

$$\text{Mean} = \mu_1' = 0$$

$\therefore$ μ_r' = rth moment about origin

$$= r\text{th moment about mean}$$
$$= \mu_r$$
$$\Rightarrow \qquad \mu_{2r+1} = \mu'_{2r+1} = 0$$

i.e., all odd order moments about mean of the standard logistic distribution are zero.

In particular $\mu_3 = 0 \Rightarrow \beta_1 = 0$

4. The mean and variance of the logistic Variable (X) with parameters α and β are given by :

$$E(X) = E(\alpha + \beta Y) \qquad \left(\because Y = \frac{X - \alpha}{\beta} \right)$$
$$= \alpha + \beta E(Y)$$
$$= \alpha$$
$$\text{Var } X = \text{Var}(\alpha + \beta Y) = \beta^2 \text{Var}(Y) = \beta^2 \pi^2/3.$$

5. We have :

$$G(y) = (1 + e^{-y})^{-1} = \left(\frac{1 + e^y}{e^y} \right)^{-1} = \frac{e^y}{1 + e^y}$$

$$\Rightarrow \qquad 1 - G(y) = 1 - \frac{e^y}{1 + e^y} = \frac{1}{1 + e^y}$$

$$\therefore G(y) . [1 - G(y)] = \frac{e^y}{(1 + e^y)^2} = g(y) \qquad \ldots(8.26j)$$

(*c.f.* Remark 3)

Also $$\frac{G(y)}{1 - G(y)} = e^y \Rightarrow y = \log_e \left[\frac{G(y)}{1 - G(y)} \right] \qquad \ldots(8.26k)$$

6. Mean deviation for the standard Logistic distribution is

$$2\left[1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots \right] = 2 \sum_{i=1}^{\infty} \left[\frac{(-1)^{i-1}}{i} \right] = 2 \log_e 2$$

Proof is left as an exercise to the reader.

EXERCISE 8(e)

1. (*a*) Show that for the exponential distribution

$$p(x) = y_0 . e^{-x/\sigma}, \; 0 \le x < \infty \; ; \; \sigma > 0,$$

mean and variance are equal. Also obtain the interquartile range of the distribution.

[Delhi Univ. B.Sc. (Stat. Hons.), 1985, 1982]

(*b*) Suppose that during rainy season on a tropical island the length of the shower has an exponential distribution, with parameter $\lambda = 2$, time being measured in minutes. What is the probability that a shower will last more than three minutes? If a shower has already lasted for 2 minutes, what is the probability that it will last for at least one more minute ? **[Madras Univ. B.Sc. (Main Stat) 1988]**

2. (*a*) If $X_1, X_2, \ldots, X_n$ are independent random variables having exponential distribution with parameter λ, obtain the distribution of $Y = \sum_{i=1}^{n} X_i$.

(*b*) Obtain the moment generating function and the cumulant generating function of the distribution with p.d.f.

$$f(x) = \frac{1}{\sigma} e^{-x/\sigma} ; 0 < x < \infty, \sigma > 0$$

[Madras Univ. B.Sc. (Main Stat.) Oct. 1992]

Hence or otherwise obtain the values of the contants $\beta_1, \beta_2, \gamma_1$ and γ_2.

(*c*) A continuous random variable X has the probability density function $f(x)$ given by

$$f(x) \begin{array}{ll} = A e^{-x/5} & x > 0 \\ = 0, & \text{otherwise} \end{array}$$

Find the value of A and show that for any two positive numbers s and t,

$$P[X > s + t \mid X > s] = P[X > t].$$

3. If X_1 and X_2 are independent and identically distributed each with frequency function $e^{-x}, x > 0$, find the frequency function of $X_1 + X_2$.

(*b*) If $X_1, X_2, \ldots, X_n$ are independent r.v.'s X_i having an exponential distribution with parameter $\theta_i, (i = 1, 2, \ldots, n)$, then prove that $Z = \min(X_1, X_2, \ldots, X_n)$ has an exponential distribution with parameter $\sum_{i=1}^{n} \theta_i$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990, '88, '86]

4. Let X and Y have common p.d.f. $\alpha e^{-\alpha x}, 0 < x < \infty, \alpha > 0$. Find the p.d.f. of

(*i*) X^3, (*ii*) $3 + 2X$, (*iii*) $X - Y$, and (*iv*) $|X - Y|$

Ans. (*i*) $\frac{\alpha}{3} x^{-2/3} \exp(-\alpha x^{1/3})$, (*ii*) $\frac{\alpha}{2} e^{-\alpha(x-3)/2}, x > 3$

(*iii*) $\frac{\alpha}{2} e^{-\alpha|x|}$, all x, and (*iv*) $\alpha e^{-\alpha x}, x > 0$.

5. (*a*) X and Y are independent random variables each exponentially distributed with the same parameter θ. find p.d.f. for $\dfrac{X}{X+Y}$ and identify its distribution. **[Delhi Univ. B.Sc. (Stat. Hons.), 1989]**

(*b*) The density functions of the independent random variables X and Y are:

$f_X(x) = \lambda e^{-\lambda x}$, $x > 0, \lambda > 0$	$f_Y(y) = \lambda e^{-\lambda y}$, $y > 0, \lambda > 0$
$= 0$, $x \le 0$	$= 0$, otherwise

Find the density function of the random variable $Z = X/Y$.

6. (*a*) For the distribution given by the density function

$$f(x) = \tfrac{1}{2} e^{-|x|}, -\infty < x < \infty,$$

obtain the moment generating function.

(*b*) Find the characteristic function of standard Laplace distribution and hence find its mean and standard deviation.

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

7. (*a*) If X has exponential distribution with mean 2, find $P(X < 1) \mid X < 2)$

Ans. $P(X < 1)$ $[P(X < 2) = (1 - e^{-\theta})/(1 - e^{2\theta})$ where $\theta = 1/2$.

[Delhi Univ. B.A. (Spl. Hons. Course-Statistics), 1989]

(*b*) If $X \sim$ Expo (λ) with $P(X \leq 1) = P(X > 1)$, ...(*)

find Var X. **[Delhi Univ. B.Sc. (Maths Hons.), 1985]**

Hint. $P(X \leq 1) + P(X > 1) = 1 \Rightarrow P(X \leq 1) = 1/2$ [Using (*)]

Ans. Var $(X) = 1/\lambda^2 = 1/(\log e^{-})^2$

8. (*a*) Show that $Y = -(1/\lambda) \log F(X)$ is Expo (λ).

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1985]

Hint.
$$\mu_Y(t) = E(e^{tY}) = E \exp\left[-\frac{t}{\lambda} \log F(x)\right]$$
$$= E[F(X)^{-t/\lambda}] = E[Z^{-t/\lambda}] \text{ where } Z = F(X) \sim U[0, 1]$$

(*b*) If X_1, X_2, X_3 and X_4 are i.i.d. $N(0, 1)$ variates, show that $Y = X_1 X_2 - X_3 X_4$, has p.d.f.

$$f(y) = \frac{1}{2} \exp[-|y|], -\infty < y <$$

[Indian Civil Services, 1984]

Hint. Show that $\varphi_Y(t) = 1/(1 + t^2) \Rightarrow Y$ has Standard Laplace distribution.

Use :
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(ax^2 + 2hxy + by^2)}\, dx\, dy = \frac{\pi}{\sqrt{ab - h^2}}$$

9. 200 electric light bulbs were tested and the average life time of the bulbs was found to be 25 hours. Using the summary given below, test the hypothesis that the lifetime is exponentially distributed.

Lifetime in hours :	0–20	20–40	40–60	60–80	80–100
Number of bulbs :	104	56	24	12	4

[You are given that an exponential distribution with parameters $\alpha > 0$ has the probability density function:

$$p(x) = \alpha e^{-\alpha x}, \quad (x \geq 0)$$
$$= 0, \quad (x < 0)$$

[Institute of Actuaries (London), April 1978]

10. Find the first four cumulants of the Laplace distribution defined by

$$f(x) = \frac{1}{2\lambda}\left[\exp\{-|x-\mu|/\lambda\}\right]; -\infty < x < \infty, \lambda > 0$$

and hence find the values of m, σ, γ_1 and γ_2. Calculate also the semi-interquartile range (S.I.R.)

Ans. $\kappa_1 = \mu$, $\kappa_2 = 2\lambda^2$, $\kappa_3 = 0$, $\kappa_4 = 12\lambda^4$; $m = \mu$, $\sigma = \sqrt{2}\,\lambda$, $\gamma_1 = 0$, $\gamma_2 = 3$ and S.I.R. $= \lambda \log_e 2$

11. The p.d.f. of a r.v. X follows the following probability law

$$p(x) = \frac{1}{2\theta}\exp\left(-\frac{|x-\theta|}{\theta}\right), \quad -\infty < x < \infty.$$

Find m.g.f. of X. Hence or otherwise, find $E(X)$ and $\text{Var}(X)$.

[Delhi Univ. B.Sc. (Stat. Hons.), 1986]

12. $X_i, i = 1, 2, ..., n$ are i.i.d. r.v.'s having Weibul distribution with three parameters. Show that the variable $Y = \min(X_1, X_2, ..., X_n)$, also has Weibul distribution and identify its parameters.

[Delhi Univ. B.Sc. (Stat. Hons.), 1984]

13. Obtain the moment generating function of Logistic distribution and hence find its mean and variance. **[Delhi Univ. B.Sc. (Stat. Hons.), 1993]**

14. (*a*) Obtain the p.d.f. $g(y)$ and the distribution function $G(y)$ of the standard logsitic variate and prove that :

(*i*) $g(y)$ is symmetric about origin.

(*ii*) $g(y) = G(y)[1 - G(y)]$

(*iii*) $y = \log_e\left[\dfrac{G(y)}{1 - G(y)}\right]$

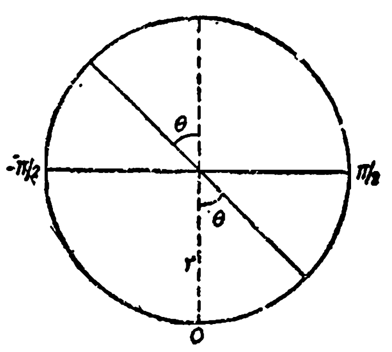
(*b*) Obtain the m.g.f. of standard logistic variate and hence prove that:

$$\text{Mean} = 0, \qquad \text{Variance} = \pi^2/3,$$

$$\beta_1 = 0, \qquad \beta_2 = \frac{21}{5}; \quad \mu_{2r+1} = 0$$

and mean deviation about mean $= 2\log_3 2$.

8·9. Cauchy's Distribution. Let us consider a roulette wheel in which the probability of the pointer stopping at any part of the circumference is constant. In other words, the probability for any value of θ lies in the interval $[-\pi/2, \pi/2]$ is constant and consequently θ is a rectangular variate in the range $[-\pi/2, \pi/2]$ with probability differential given by



$$\left.\begin{aligned} dP(\theta) &= (1/\pi)\, d\theta, -\pi/2 \le \theta \le \pi/2 \\ &= 0, \text{ otherwise} \end{aligned}\right\} \quad ...(8·27)$$

Let us now transform to the variable X by the substitution:

$x = r\tan\theta \Rightarrow dx = r\sec^2\theta\, d\theta$

Since $-\pi/2 \le \theta \le \pi/2$, the range for X is from $-\infty$ to ∞. Thus the probability differential of X becomes:

$$dF(x) = \frac{1}{\pi}\cdot\frac{dx}{r\sec^2\theta} = \frac{1}{\pi}\cdot\frac{dx}{[r\{1+(x^2/r^2)\}]} = \frac{r}{\pi}\cdot\frac{dx}{r^2+x^2}\ ;\ -\infty < x < \infty$$

In particular if we take $r = 1$, we get

$$f(x) = \frac{1}{\pi}\cdot\frac{1}{1+x^2},\ -\infty < x < \infty$$

This is the p.d.f. of a standard Cauchy variate and we write $X \sim C(1, 0)$

Definition. *A random variable X is said to have a standard Cauchy distribution if its p.d.f. is given by*

$$f_X(x) = \frac{1}{\pi(1+x^2)},\ -\infty < x < \infty \qquad \text{...(8·28)}$$

and X is termed as a standard Cauchy variate.

More generally, Cauchy distribution with parameters λ and μ has the following p.d.f.,

$$g_Y(y) = \frac{\lambda}{\pi[\lambda^2 + (y-\mu)^2]},\ -\infty < y < \infty\ ;\ \lambda > 0 \qquad \text{...(8·29)}$$

and we write $X \sim C(\lambda, \mu)$

But putting $X = (Y-\mu)/\lambda$ in (8·29), we get (8·28).

8·9·1. Characteristic Function of Cauchy Distribution. If X is a standard Cauchy variate then

$$\varphi_X(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{e^{itx}}{1+x^2}\,dx \qquad \text{...(*)}$$

To evaluate (*) consider Lapalce distribution

$$f_1(z) = \tfrac{1}{2}e^{-|z|},\ -\infty < z < \infty$$

Then $\qquad \varphi_1(t) = E(e^{itZ}) = \dfrac{1}{1+t^2}$ [From (8·25 a)

Since $\varphi_1(t)$ is absolutely integrable in $(-\infty, \infty)$, we have by Inversion theorem

$$\frac{1}{2}e^{-|z|} = f_1(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itz}\varphi_1(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-itz}}{1+t^2}\,dt$$

$$\Rightarrow \qquad e^{-|z|} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{e^{-itz}}{1+t^2}\,dt = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{e^{itz}}{1+t^2}\,dt \qquad \text{(Changing } t \text{ to } -t)$$

On interchanging t and z, we get

$$e^{-|t|} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{itz}}{1+z^2}\, dz \qquad \ldots(**)$$

From (*) and (**), we get

$$\varphi_X(t) = e^{-|t|} \qquad \ldots(8\cdot30)$$

Remarks. 1. If Y is a Cauchy variate with parameters λ and μ, then

$$X = \frac{Y-\mu}{\lambda} \quad \Rightarrow \quad Y = \mu + \lambda X$$

$$\therefore \qquad \varphi_Y(t) = E(e^{itY}) = e^{i\mu t} E(e^{it\lambda X}) = e^{i\mu t} \varphi_X(t\lambda)$$

$$= e^{i\mu t - \lambda |t|}, \lambda > 0 \qquad \ldots(8\cdot30\, a)$$

2. Additive Property of Cauchy distribution. *If* X_1 and X_2 *are independent Cauchy variates with parameters* (λ_1, μ_1) and (λ_2, μ_2) *respectively, then* $X_1 + X_2$ *is a Cauchy variate with parameters* $(\lambda_2 + \lambda_2, \mu_1 + \mu_2)$.

Proof. $\varphi_{X_j}(t) = \exp\{ i\mu_j t - \lambda_j |t| \}, (j = 1, 2)$

$$\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t)\, \varphi_{X_2}(t) \qquad \text{(Since } X_1, X_2 \text{ are independent)}$$

$$= \exp\left[it(\mu_1 + \mu_2) - (\lambda_1 + \lambda_2)|t| \right]$$

and the result follows by uniqueness theorem of characteristic functions.

3. Since $\varphi'_X(t)$ in (8·30) [where (′) denotes differentiation w.r.t. t] does not exist at $t = 0$, the mean of the Cauchy distribution does not exist.

4. Let $X_1, X_2, \ldots, X_n$ be a sample of n independent observations from a standard Cauchy distribution and define $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then

$$\varphi_{\bar{X}}(t) = \varphi_{\Sigma X_i}(t/n) = \prod_{i=1}^{n} [\varphi_{X_i}(t/n)]$$

$$= [\varphi_{X_i}(t/n)]^n \qquad \text{(since } X_i\text{'s are } i.i.d.\text{)}$$

$$= [e^{-|t/n|}]^n = e^{-|t|} = \varphi_X(t)$$

Hence by uniqueness theorem of characteristic functions, we have:

"*The arithmetic mean* $\bar{X}$ *of a sample* $X_1, X_2, \ldots, X_n$ *of independent observations from a standard Cauchy distribution is also a standard Cauchy variate.* In other words, the arithmetic mean of a random sample of any size yields exactly as much information as a single determination of X."

This implies that the sample mean $\bar{X}_n$ of a random sample of size n, as an estimate of population mean does not improve with increasing n, which contradicts the Weak **Law of Large Numbers (WLLN).**

8 9 2. Moments of Cauchy Distribution.

$$E(Y) = \int_{-\infty}^{\infty} y f(y)\, dy = \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{y}{\lambda^2 + (y-\mu)^2}\, dy$$

$$= \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{(y-\mu)+\mu}{\lambda^2+(y-\mu)^2}\, dy$$

$$= \mu \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{dy}{\lambda^2+(y-\mu)^2} + \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{(y-\mu)}{\lambda^2+(y-\mu)^2}\, dy$$

$$= \mu \,.\, 1 + \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{z}{\lambda^2+z^2}\, dz$$

Although the integral $\int_{-\infty}^{\infty} \frac{z}{\lambda^2+z^2}\, dz$, is not completey convergent, *i.e.*, $\lim\limits_{\substack{n \to \infty \\ n' \to \infty}} \int_{-n}^{n'} \frac{z}{\lambda^2+z^2}\, dz$ does not exist, its principal value, *viz.*, $\lim\limits_{n \to \infty} \int_{-n}^{n} \frac{z}{\lambda^2+z^2}\, dz$ exists and is equal to zero. Thus, in the general sense the mean of Cauchy distribution does not exist. But, if we conventionally agree to assume that the mean of Cauchy distribution exists (by taking the principal value), then it is located at $x = \mu$. Also, obviously, the probability curve is symmetrical about the point $x = \mu$. Hence for this distribution, the mean, median and mode coincide at the point $x = \mu$.

$$\mu_2 = E(Y-\mu)^2 = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\, dy = \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{(y-\mu)^2}{\lambda^2+(y-\mu)^2}\, dy,$$

which does not exist since the integral is not convergent. Thus, in general, for the Cauchy's distribution μ_r, $(r \geq 2)$ do not exist.

Remark. The role of Cauchy distribution in statistical theory often lies in providing counter examples, *e.g.* it, is often quoted as a distribution for which moments do not exist. It also provides an example to show that

$$\varphi_{X+Y}(t) = \varphi_X(t)\, \varphi_Y(t)$$

does *not* imply that X and Y are independent. [See Remark to Theorem 6·23]

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a standard Cauchy distribution. Let $\bar{X} = \sum_{i=1}^{n} X_i / n$. Since $E(X_i)$ does not exist ($\because$ mean of a Cauchy distribution does not exist), $E(\bar{X})$ does not exist either and the definition of an unbaised estiamte does not apply to $\bar{X}$.

Cauchy distribution also contradicts the WLLN [See Remark 4, § 8·9·1].

Example 8·42. *Let X have a (standard) Cauchy distribution. Find a p.d.f. for X^2 and identify its distribution.* **[Delhi Univ. B.Sc. (Stat. Hons.), 1989; '87]**

Solution. Since X has a standard Cauchy distribution, its p.d.f. is

$$f(x)=\frac{1}{\pi}\;\frac{1}{1+x^2},\ -\infty<x<\infty$$

The distribution function of $Y=X^2$ is

$$G_Y(y)=P(Y\le y)=P(X^2\le y)=P(-\sqrt{y}\le X\le\sqrt{y})$$

$$=\int_{-\sqrt{y}}^{\sqrt{y}} f(x)\,dx=2\,\frac{1}{\pi}\int_0^{\sqrt{y}}\frac{dx}{1+x^2}$$

$$=\frac{2}{\pi}\tan^{-1}(\sqrt{y}),\ 0<y<\infty$$

The p.d.f. $g_Y(y)$ of Y is given by

$$g_Y(y)=\frac{d}{dy}\left[G_Y(y)\right]=\frac{2}{\pi}\cdot\frac{1}{(1+y)}\cdot\frac{1}{2\sqrt{y}}$$

$$=\frac{1}{\pi}\cdot\frac{y^{-1/2}}{1+y}=\frac{1}{B\left(\frac{1}{2},\frac{1}{2}\right)}\cdot\frac{y^{(1/2)-1}}{(1+y)^{(1/2+1/2)}},\ y>0$$

This is the p.d.f. of Beta distribution of second kind with parameters $\left(\frac{1}{2},\frac{1}{2}\right)$, *i.e.*, $X^2\sim\beta_2\left(\frac{1}{2},\frac{1}{2}\right)$

Remark. Here $y=g(x)=x^2$, gives $g'(x)=2x$ which is sometimes >0 and sometimes <0. Hence Theorem 5·9 can not be used in this case.

Example 8·43. *Let $X\sim N(0,1)$ and $Y\sim N(0,1)$ be independent random variables. Find the distribution of X/Y and identify it.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1990; Nagpur Univ. B.Sc., 1991]

Solution. Since X and Y are independent $N(0,1)$, their joint p.d.f. is given by

$$f_{XY}(x,y)=f_X(x)\cdot f_Y(y)=\frac{1}{2\pi}\cdot e^{-(x^2+y^2)/2}$$

Let us make the following transformation of variables

$$u=x/y,\ v=y\ \text{ so that }\ x=uv,\ y=v$$

Jacobian of transformation $J=v$.

Hence the joint p.d.f. of U and V becomes

$$g_{UV}(u,v)=\frac{1}{2\pi}\cdot\exp\left\{-(u^2v^2+v^2)/2\right\}|J|$$

$$=\frac{1}{2\pi}\exp\left\{-(1+u^2)\,v^2/2\right\}|v|,\ -\infty<(u,v)<\infty$$

The marginal p.d.f. of U is

$$g_U(u)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\left\{-(1+u^2)\,v^2/2\right\}|v|\,dv$$

$$=\frac{1}{\pi}\int_0^{\infty}e^{-t}\frac{dt}{(1+u^2)}\qquad\left[\left(\tfrac{1}{2}(1+u^2)\,v^2=t\right)\right]$$

$$=\frac{1}{\pi(1+u^2)}\Big|-e^{-t}\Big|_0^{\infty}=\frac{1}{\pi(1+u^2)},\ -\infty<u<\infty$$

which is the p.d.f. of a standard Cauchy distribution.

Thus the ratio of two independent standard normal variates is a standard Cauchy variate.

Example 8·44. *Let X and Y be i.i.d. standard Cauchy variates. Prove that the p.d.f. of XY is* : $\dfrac{2}{\pi^2}\left\{\dfrac{\log|x|}{x^2-1}\right\}$. **[Delhi Univ. M.Sc. (Stat), 1991]**

Solution. Since X and Y are independent standard Cauchy variates, their joint p.d.f. is given by

$$f(x,y)=\frac{1}{\pi^2}\cdot\frac{1}{(1+x^2)(1+y^2)}\ ;-\infty<(x,y)<\infty\ .$$

Let $u=xy$ and $v=y$. Then Jacobian of transformation is given by

$$J=\frac{\partial(x,y)}{\partial(u,v)}=\begin{vmatrix}\dfrac{1}{v} & -\dfrac{u}{v^2}\\ 0 & 1\end{vmatrix}=\frac{1}{v}\qquad\left(\because y=v,\ x=\frac{u}{v}\right)$$

Thus the joint p.d.f. of U and V is given by

$$g(u,v)=\frac{1}{\pi^2}\frac{1}{\left(1+\dfrac{u^2}{v^2}\right)(1+v^2)}\cdot\frac{1}{|v|}$$

$$=\frac{1}{\pi^2}\frac{|v|}{(u^2+v^2)(1+v^2)}\ ;-\infty<(u,v)<\infty$$

Integrating w.r.to v over the range $-\infty$ to ∞, the marginal p.d.f. of U is given by

$$g_1(u)=\int_{-\infty}^{\infty}g(u,v)\,dv=\frac{1}{\pi^2}\int_{-\infty}^{\infty}\frac{|v|}{(u^2+v^2)(1+v^2)}\,dv$$

$$=\frac{2}{\pi^2}\int_0^{\infty}\frac{|v|\,dv}{(u^2+v^2)(1+v^2)},$$

(Since the integrand is an even function of v.)

$$\therefore\quad g_1(u)=\frac{2}{\pi^2}\int_0^{\infty}\frac{v}{(u^2+v^2)(1+v^2)}\,dv$$

$$=\frac{1}{\pi^2}\int_0^{\infty}\frac{2v}{(u^2-1)}\left(\frac{1}{1+v^2}-\frac{1}{u^2+v^2}\right)dv$$

$$=\frac{1}{\pi^2(u^2-1)}\Big|\log(1+v^2)-\log(u^2+v^2)\Big|_0^{\infty}$$

$$=\frac{1}{\pi^2(u^2-1)}\left|\log\left(\frac{1+v^2}{u^2+v^2}\right)\right|_0^{\infty}$$

$$= \frac{1}{\pi^2 (u^2 - 1)} \left\{ \left[\left(\log \left(\frac{\frac{1}{v^2} + 1}{\frac{u^2}{v^2} + 1} \right) \right) \right]_{v = \infty} - \log \left(\frac{1}{u^2} \right) \right\}$$

$$= \frac{1}{\pi^2 (u^2 - 1)} [\log 1 + 2 \log |u|] = \frac{2 \log |u|}{\pi^2 (u^2 - 1)}, -\infty < u < \infty$$

EXERCISE 8(f)

1. (*a*) Show that a function

$$f(x; \mu, \lambda) = \frac{k}{\lambda^2 + (x - \mu)^2}; -\infty < x < \infty$$

represents a frequency function of a distribution for a suitable value of k. Determine k and obtain median and quartiles of the distribution. Hence interpret the parameters λ and μ of the distribution.

(*b*) If X is a Cauchy variate with parameters λ and μ, find the characteristic function $\varphi_X(t)$. Discuss briefly the role of Cauchy distribution in Statistics.

[Bombay Univ. B.Sc. (Stat.), 1993]

(*c*) "The role of Cauchy distribution often lies in providing counter examples." Justify. **[Delhi Univ. B.Sc. (Stat. Hons.), 1991, '88]**

(*d*) Discuss briefly the role of Cauchy distribution in statistics.

If $X_1, X_2, ..., X_n$ are independent standard Cauchy variates, show that the mean $\bar{X} = (X_1 + X_2 + ... + X_n)/n$, is also a Cauchy variate.

[Delhi Univ. B.Sc. (Stat. Hons.), 1986]

2. (*a*) If X and Y are independent random variables following Cauchy distribution with parameters (λ_1, μ_1) and (λ_2, μ_2) respectively, show that $X + Y$ follows Cauchy distribution with parameters $\lambda_1 + \lambda_2$ and $\mu_1 + \mu_2$.

(*b*) Obtain the characteristic function of Cauchy distribution

$$dF(x) = \frac{dx}{\pi (1 + x^2)}, -\infty < x < \infty$$

If $X_1, X_2, ..., X_n$ are independent Cauchy variates, show that the mean $\bar{X} = \frac{1}{n} \Sigma X$ is also a Cauchy variate.

3. Let X and Y be standard normal variates. Find the distribution of $U = X/|Y|$.

Ans. $$f(u) = \frac{1}{\pi} \cdot \frac{1}{1 + u^2}, -\infty < u < \infty$$

4. A needle spins about the point $(0, b)$ of the x-y plane with $b > 0$ and comes to a stop thereby making an angle φ with Y-axis. The direction of the needle then intersects the x-axis at a point $(X, 0)$. Assuming φ is a r.v. with uniform

probability distribution on $(-\pi/2, \pi/2)$, what is the distribution function and hence p.d.f. of X ?

Ans. $F_X(x) = \frac{1}{\pi}\left[\tan^{-1}(x/b) + \frac{\pi}{2}\right], f_X(x) = \frac{1}{\pi}\cdot\frac{b}{x^2 + !^2}, -\infty < X < \infty$

5. If X_1, X_2, X_3, X_4 are independent standard normal variates, find the distribution of $\frac{X_1}{X_2} + \frac{X_3}{X_4}$.

6. $\bar{X}_n$ is the mean of n independent random variables distributed like X, and X has a symmetric distribution. If $\bar{X}_n$ has exactly the same distribution as X for all n, then prove that the characteristic function of X is

$$\Phi_X(t) = e^{-c|t|}$$

for some real constant $c > 0$. Identify this distribution.

7. If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent random varaibles, obtain the p.d.f. of $U = \frac{X - \mu_1}{Y - \mu_2}$. (I.I.T., B. Tech. 1993)

8·10. Central Limit Theorem. The central limit theorem in the mathematical theory of probability may be expressed as follows :

"*If X_i, $(i = 1, 2, ..., n)$ be independent random variables such that $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, then it can be proved that under certain very general conditions, the random variable $S_n = X_1 + X_2 + \ldots + X_n$, is asymptotically normal with mean μ and standard deviation σ where*

$$\mu = \sum_{i=1}^{n} \mu_i \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2$$

This theorem was first stated by Laplace in 1812 and a regorous proof under fairly general conditions was given by Liapounoff in 1901. Below we shall consider some particular cases of this general central limit theorem.

De-Moivre's–Laplace theorem. (1733). A particular case of central limit theorem is De-Moivre's theorem which states as follows:

"If $$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q \end{cases}$$

then the distribution of the random variable $S_n = X_1 + X_2 + \ldots + X_n$, where X_i's are independent, is asymptotically normal as $n \to \infty$."

Proof. M.G.F. of X_i is given by

$$M_{X_i}(t) = E(e^{tX_i}) = e^{t\cdot 1}p + e^{t\cdot 0}q = (q + pe^t)$$

M.G.F. of the sum $S_n = X_1 = X_2 + \ldots + X_n$ is given by

$$M_{S_n}(t) = M_{X_1 + X_2 + \ldots + X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \ldots M_{X_n}(t)$$

$$= [M_{X_i}(t)]^n \qquad \text{(since } X_i\text{'s are identically distributed)}$$

$$= (q + pe^t)^n,$$

which is the M.G.F. of a binomial variate with parameters n and p.

$\therefore \qquad E(S_n) = np = \mu$ (say), and $V(S_n) = npq = \sigma^2$, (say).

Let
$$Z \doteq \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - \mu}{\sigma}$$

$$M_Z(t) = e^{-\mu t/\sigma} M_{S_n}(t/\sigma) \qquad \text{[c.f. Chapter 6]}$$
$$= e^{-npt/\sqrt{npq}} \left[q + pe^{t/\sqrt{npq}} \right]^n$$
$$= \left[1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n \qquad \text{[c.f. Example 7·19]}$$

where $O(n^{-3/2})$ represents terms involving $n^{3/2}$ and higher powers of n in the denominator.

Proceeding to the limits as $n \to \infty$, we get

$$\lim_{n \to \infty} M_Z(t) = \lim_{n \to \infty} \left[1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n = \lim_{n \to \infty} \left[1 + \frac{t^2}{2n} \right]^n = e^{t^2/2}$$

which is the M.G.F. of a standard normal variate. Hence by the uniqueness theorem of M.G.F.'s

$$Z = \frac{S_n - \mu}{\sigma} \text{ is asymptotically } N(0, 1).$$

Hence $S_n = X_1 + X_2 + \ldots + X_n$ is asymptotically $N(\mu, \sigma^2)$ as $n \to \infty$.

Remarks 1. From this theorem it follows that standard binomial variate tends to standard normal variate as $n \to \infty$. In other words, binomial distribution tends to normal distribution as $n \to \infty$.

2. Convergence in Distribution or Law. Let $\{X_n\}$ be a sequence of r.v.'s and $\{F_n\}$ be the corresponding sequence of distribution functions. We say that X_n converges in distribution (or law) to X if there exists a r.v. X with distribution function F such that as $n \to \infty$, $F_n(x) \to F(x)$ at every point x at which F is continuous.

$$\text{We write } X_n \xrightarrow{L} X \text{ or } X_n \xrightarrow{d} X.$$

3. It may be remarked that convergence in probability discussed in § 6·14 implies convergence in distribution (or law) *i.e.*,

$$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X \qquad \ldots(*)$$

The converse is not true i.e., $X_n \xrightarrow{L} X$, in general, does not imply $X_n \xrightarrow{P} X$

However, we have the following result.

Let k be a constant. Then

$$X_n \xrightarrow{L} k \Rightarrow X_n \xrightarrow{P} k \qquad \ldots(**)$$

Combining (*) and (**), we get the following result.

Let k be a constant. Then

$$X_n \xrightarrow{L} k \Leftrightarrow X_n \xrightarrow{P} k \qquad \ldots(***)$$

8·10·1 Lindeberg-Levy Theorem. The following case of central limit theorem for equal components, *i.e.*, for identically distributed variables, was first proved by Lindeberg and Levy.

"If $X_1, X_2, ..., X_n$ are independently and identically distributed random variables with

$$\left.\begin{aligned} E(X_i) &= \mu_1 \\ V(X_i) &= \sigma_1^2 \end{aligned}\right\} i = 1, 2, ..., n$$

then the sum $S_n = X_1 + X_2 + ... + X_n$ is asymptotically normal with mean $\mu = n\mu_1$ and variance $\sigma^2 = n\sigma_1^2$."

Here we make the following assumptions :

(*i*) The variables are independent and identically distributed

(*ii*) $E(X_i^2)$ exists for all $i = 1, 2, ...$

Proof. Let $M_1(t)$ denote the M.G.F. of each of the deviation $(X_i - \mu_1)$ and $M(t)$ denote the M.G.F. of the standard variate

$$Z = (S_n - \mu)/\sigma$$

Since μ_1' and μ_2', (about origin) of the deviation $(X_i - \mu_1)$ are given by

$$\mu_1' = E(X_i - \mu_1) = 0,\ \mu_2' = E(X_i - \mu_1)^2 = \sigma_1^2$$

We have

$$M_1(t) = \left(1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + ...\right)$$

$$= \left[1 + \frac{t^2}{2!}\sigma_1^2 + O(t^3)\right] \qquad ...(*)$$

where $O(t^3)$ contains terms with t^3 and higher powers of t.

We have

$$Z = \frac{S_n - \mu}{\sigma} = \frac{(X_1 + X_2 + ... + X_n) - n\mu_1}{\sigma} = \sum_{i=1}^{n} \left(\frac{X_i - \mu_1}{\sigma}\right)$$

and since X_i's are independent, we get

$$M_Z(t) = M_{\sum_{i=1}^{n} (X_i - \mu_1)/\sigma_1}(t) = M_{\sum_{i=1}^{n} (X_i - \mu_1)}(t/\sigma)$$

$$= \prod_{i=1}^{n} \left\{M_{(X_i - \mu_1)}(t/\sigma)\right\} = [M_1(t/\sigma)]^n = [M_1(t/\sqrt{n}\,\sigma_1)]^n$$

$$= \left[1 + \frac{t^2}{2n} + O(n^{-3/2})\right]^n \qquad \text{[From (*)]}$$

For every fixed 't', the terms $O(n^{-3/2}) \to 0$ as $n \to \infty$. Therefore, as $n \to \infty$, we get

$$\lim_{n \to \infty} M_Z(t) = \lim_{n \to \infty} \left[1 + \frac{t^2}{2n} + O(n^{-3/2})\right]^n = \exp\,[t^2/2],$$

which is the M.G.F. of standard normal variate. Hence by uniqueness theorem of M.G.F.'s $Z = (S_n - \mu)/\sigma$ is asymptotically $N(0, 1)$, or $S_n = X_1 + X_2 + \ldots + X_n$ is asymptotically $N(\mu, \sigma^2)$, where $\mu = n\mu_1$ and $\sigma^2 = n\sigma_1^2$.

Note. *C.L.T.* can be stated in another form as follows:

(*i*) If $X_1, X_2, \ldots, X_n$ are *i.i.d.* with mean μ_1 and variance σ_1^2 (finite) and $S_n = X_1 + X_2 + \ldots + X_n$, then

$$\lim_{n \to \infty} P\left[a \leq \frac{S_n - n\mu_1}{\sigma_1\sqrt{n}} \leq b\right] = \Phi(b) - \Phi(a)$$

$$= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \qquad \ldots(8\cdot31)$$

for $-\infty < a < b < \infty$; $\Phi(-\infty) = 0, \Phi(\infty) = 1$

or

$$(ii) \quad \lim_{n \to \infty} P\left[a \leq \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \leq b\right] = \Phi(b) - \Phi(a) \qquad \ldots(8\cdot31\cdot1)$$

or, still another form :

$$(iii) \quad \lim_{n \to \infty} P\left[a \leq \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}} \leq b\right] = \Phi(b) - \Phi(a)$$

$$i.e., \quad \lim_{n \to \infty} P\left[a \leq \frac{\bar{X}_n - \mu_1}{\sigma_1/\sqrt{n}} \leq b\right] = \Phi(b) - \Phi(a) \qquad \ldots(8\cdot31\cdot2)$$

Remarks 1. We wrote the *C.L.T.* using non-strict inequalities

$$P[a \leq (.) \leq b]$$

It makes no difference whether one or both are changed to a strict inequality. The reason is that the limit distribution function (*d.f.*) $\Phi(.)$ is a continuous *d.f.*

2. In the binomial case, C.L.T. gives good approximation if p is nearly 1/2. For p near about 0 or 1, the C.L.T. approximation still holds but in that case n has to be sufficiently greater than in the case $p = 1/2$ approximately.

8·10·2. Applications of Central Limit Theorem. (*a*) If $X_1, X_2, \ldots$ are *i.i.d.* $B(r, p)$ and

$$S_n = X_1 + X_2 + \ldots + X_n, \text{ then}$$

$$E(S_n) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} (rp) = nrp$$

and

$$V(S_n) = V\left(\sum_{1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} (rpq) = nrpq$$

Hence (8·31·1)

$$\Rightarrow \lim_{n \to \infty} P\left[a \leq \frac{S_n - nrp}{\sqrt{nrp(1-p)}} \leq b\right] = \Phi(b) - \Phi(a), 0 < p < 1$$

(*b*) If Y_n is binomial variate with parameters n and p, then

$$\lim_{n \to \infty} P\left[a \le \frac{Y_n - np}{\sqrt{np(1-p)}} \le b\right] = \Phi(b) - \Phi(a), 0 < p < 1$$

Proof. Let $X_1, X_2, \ldots$ be *i.i.d.* Bernoulli variates, *i.e.*, $B(1, p)$, then
$S_n = X_1 + X_2 + \ldots + X_n = B(n, p)$. But $Y_n = B(n, p)$

Hence using Y_n instead of S_n in (8·31·1), we get

$$\lim_{n \to \infty} P\left[a \le \frac{Y_n - E(Y_n)}{\sqrt{Var\, Y_n}} \le b\right] = \Phi(b) - \Phi(a)$$

i.e.,
$$\lim_{n \to \infty} P\left[a \le \frac{Y_n - np}{\sqrt{npq}} \le b\right] = \Phi(b) - \Phi(a), q = 1 - p$$

(*c*) If Y_n is distributed as $P(n)$, then

$$\lim_{n \to \infty} P\left[a \le \frac{Y_n - n}{\sqrt{n}} \le b\right] = \Phi(b) - \Phi(a)$$

Thus, for instance

$$\lim_{n \to \infty} P\left(Y_n \le n\right) = \frac{1}{2}, \textit{ i.e., } \sum_{k=0}^{n} \frac{e^{-n} n^k}{k!} = \frac{1}{2} \textit{ as } n \to \infty$$

Proof. Let $X_1, X_2, \ldots$ be *i.i.d.* $P(1)$. Then
$S_n = X_1 + X_2 + \ldots + X_n \sim P(n) \Rightarrow Y_n = S_n$

$$\therefore \quad P\left[a \le \frac{Y_n - n}{\sqrt{n}} \le b\right] = P\left[a \le \frac{S_n - n}{\sqrt{n}} \le b\right]$$
$$\to \Phi(b) - \Phi(a) \text{ as } n \to \infty$$

In particular, let us take $a = -\infty$ and $b = 0$, then

$$P\left(a \le \frac{Y_n - n}{\sqrt{n}} \le b\right) = P\left(\frac{Y_n - n}{\sqrt{n}} \le 0\right) = P(Y_n \le n) \qquad \ldots(*)$$

Also $\Phi(b) - \Phi(a) = \Phi(0) - \Phi(-\infty) = 1/2 \qquad \ldots(**)$

From (*) and (**), we get

$P(Y_n \le n) \to 1/2$ as $n \to \infty$

Remark. This result could be generalised. In fact, on taking $a = -\infty$ and $b = 0$ in (8·31·1), we get

$$P\left[\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \le 0\right] \to 1/2 \quad \Rightarrow \quad P[S_n \le E(S_n)] \to 1/2 \text{ as } n \to \infty$$

8·10·3. Liapounoff's Central Limit Theorem. Below we shall give (without proof) the central limit theorem for the generalised case when the variables are not identically distributed and where, in addition to the existence of the second moment for the variables X_i, we impose some further conditions.

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that

$$\left.\begin{array}{l} E(X_i) = \mu_i \\ V(X_i) = \sigma_i^2 \end{array}\right\}; i = 1, 2, \ldots, n.$$

Let us suppose that third absolute moment of X_i about its mean *viz.*,

$$\rho_i^3 = E\left\{ |X_i - \mu_i|^3 \right\} ; i = 1, 2, ..., n$$

is finite. Further let

$$\rho^3 = \sum_{i=1}^{n} \rho_i^3$$

If $\lim\limits_{n \to \infty} \dfrac{\rho}{\sigma} = 0$, the sum $X = X_1 + X_2 + \dots + X_n$ is asymptotically $N(\mu, \sigma^2)$, where $\mu = \sum\limits_{i=1}^{n} \mu_i$ and $\sigma^2 = \sum\limits_{i=1}^{n} \sigma_i^2$.

Remarks. 1. *(About Liapounoff's theorem).* If the variables $X_i ; i = 1, 2, ..., n$ are identical, then

$$\rho^3 = \sum_{i=1}^{n} \rho_i^3 = n \rho_1^3 \text{ and } \sigma^2 = \sum_{i=1}^{n} \sigma_i^2 = n \sigma_1^2$$

$$\therefore \qquad \frac{\rho}{\sigma} = \frac{n^{1/3} \rho_1}{n^{1/2} \cdot \sigma_1} = \frac{\rho_1}{\sigma_1} \cdot \frac{1}{n^{1/6}} \to 0 \text{ as } n \to \infty .$$

Thus for identical variables, the condition of Liapounoff's theorem is satisfied.

It may be pointed out here that Lindeberg - Levy theorem proved in § 8·10·1, should not be inferred as a particular case of Liapounoff's theorem, since the former does not assume the existence of the third moment.

2. Central limit theorem can be expected in the following cases:

(*i*) If a certain random variable X arises as cumulative effect of several independent causes each of which can be considered as a continuous random variable, then X obeys central limit theorem under certain regularity conditions.

(*ii*) If $\varphi(X_1, X_2, ..., X_n)$, is a function of X_i 's having first and second continuous derivatives about the point $(\mu_1, \mu_2, ..., \mu_n)$, then under certain regularity conditions, $\varphi(X_1, X_2, ..., X_n)$ is asymptotically normal with mean $\varphi(\mu_1, \mu_2, ..., \mu_n)$.

(*iii*) Under certain conditions, the central limit theorem holds for variables which are not independent.

3. Relation between CLT and WLLN. (a). Both the central limit theorem (CLT) and the weak law of large numbers hold for a sequence $\{X_n\}$ of i.i.d. random variables with finite mean μ and variance σ^2.

However, in this case the CLT is a stronger result than the WLLN in the sense that the former provides an estimate of the $P\left[\left| S_n - n\mu \right| / n \geq \varepsilon \right]$, as given below:

$$P\left[\left| \frac{S_n - n\mu}{n} \right| \geq \varepsilon \right] = P\left[\left| \bar{X}_n - \mu \right| \geq \varepsilon \right]$$

$$= P\left[\left| \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right| \geq \frac{\varepsilon}{\sigma/\sqrt{n}} \right]$$

$$= P\left[|Z| \geq \varepsilon \sqrt{n}/\sigma \right] ; Z \sim N(0, 1)$$

$$= 1 - P[\,|Z| \le \varepsilon\sqrt{n}/\sigma\,]$$

$$= 1 - \left[\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right]$$

where $\Phi(\,.\,)$ is the distribution function of standard normal variate.

However, WLLN does not require the existence of variance (c.f. Khinchin'es theorem].

(*b*) For the sequence $\{X_n\}$ of independent and uniformly bounded r.v.'s, WLLN holds [c.f. theorem 6·32] and CLT holds in this case provided

$$B_n = Var\left(X_1 + X_2 + \ldots + X_n\right) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2 \to \infty \text{ as } n \to \infty.$$

(*c*) For the sequence $\{X_n\}$ of independent r.v.'s, CLT may hold but the WLLN may not hold.

8·10·4. Cramer's Theorem. We state below (without proof), a useful result on the convergense of sequences of r.v.'s.

Cramer's Theorem. Let $\{X_n\}$ and $\{Y_n\}$ be sequences of r.v.'s such that:

$$X_n \xrightarrow{L} X \text{ and } Y_n \xrightarrow{P} c \text{ (constant)},$$

then $\dfrac{X_n}{Y_n} \xrightarrow{L} \dfrac{X}{c}$ if $c \ne 0$

For illustrations, see Example 8·46 and Qns. 15 to 17 in Exercise 8 (*g*).

Example 8·45. *Let $X_1, X_2, \ldots$ be a i.i.d. Poisson variates with parameter λ. Use CLT to estimate $P(120 \le S_n \le 160)$, where*

$$S_n = X_1 + X_2 + \ldots + X_n\,;\ \lambda = 2 \text{ and } n = 75.$$

Solution. Since X_i is *i.i.d.* $P(\lambda)$,

$$E(X_i) = \lambda \text{ and } Var(X_i) = \lambda\,;\ i = 1, 2, \ldots, n$$

$$\therefore E(S_n) = \sum_{i=1}^{n} E(X_i) = n\lambda$$

$$Var(S_n) = Var(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} Var\,X_i = n\lambda$$

Hence by Lindeberg – Levy CLT, (for large n)

$$S_n \sim N(n\lambda, n\lambda) = N(\mu = 150, \sigma^2 = 150)\,;\ (n = 75\,;\ \lambda = 2)$$

$$\therefore\ P(120 \le S_n \le 160) = P\left(\frac{120 - 150}{\sqrt{150}} \le Z \le \frac{160 - 150}{\sqrt{150}}\right)$$

$$= P(-2\cdot45 \le Z \le 0\cdot82)\,;\ Z \sim N(0, 1)$$

$$= P(-2\cdot45 \le Z \le 0) + P(0 \le Z \le 0\cdot82)$$

$$= P(0 \le Z \le 2\cdot45) + P(0 \le Z \le 0\cdot82)$$

Example 8·46. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. standardised variates with $E(X_i^4) < \infty$. Find the limiting distribution of:*

$$Z_n = \sqrt{n}\left[X_1X_2 + X_3X_4 + \ldots + X_{2n-1}X_{2n}\right] \div \left[X_1^2 + X_2^2 + \ldots + X_{2n}^2\right]$$

Solution. Since X_i s are *i.i.d.* standardised variates we have:

$$E(X_i) = 0\ ;\quad Var(X_i) = E(X_i^2) = 1,\quad i = 1, 2, ..., n \qquad ...(*)$$

Let $\qquad Y_i = X_{2i-1}X_{2i}\ ;\quad i = 1, 2, ..., n$

$\Rightarrow \qquad E(Y_i) = E(X_{2i-1})\,E(X_{2i}) = 0 \qquad$ ($\because X_i$'s are independent)

$\therefore \qquad Var(Y_i) = E\,Y_i^2 = E\left(X_{2i-1}^2 X_{2i}^2\right) = E\left(X_{2i-1}^2\right)E\left(X_{2i}^2\right) = 1$

Hence Y_i, $i = 1, 2, ..., n$ are also *i.i.d.* standardised variates. Hence by CLT for *i.i.d. r.v.'s*, $\left[S_n = \sum_{i=1}^{n} Y_i\right]$, we get

$$U_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} = \frac{X_1X_2 + X_3X_4 + \ldots + X_{2n-1}X_{2n}}{\sqrt{n}} \xrightarrow{L} Z \sim N(0, 1)$$

as $n \to \infty$

Also $E(X_i^2) = 1$ (finite), $i = 1, 2, ..., n$.

Hence by Khinchine's theorem, WLLN applies to the sequence $\{X_i^2\}$, $i = 1, 2, \ldots 2n$; so that

$$V_n = \frac{X_1^2 + X_2^2 + \ldots + X_{2n}^2}{2n} \xrightarrow{P} E(X_i^2) = 1, \text{ as } n \to \infty.$$

Hence by Cramer's theorem

$$\lim_{n \to \infty} \frac{Un}{Vn} = \frac{2\sqrt{n}\,[X_1X_2 + X_3X_4 + \ldots + X_{2n-1}X_{2n}]}{X_1^2 + X_2^2 + \ldots + X_{2n}^2} \xrightarrow{L} \frac{Z}{1} \sim N(0, 1)$$

$$\Rightarrow \lim_{n \to \infty} \frac{\sqrt{n}\,[X_1X_2 + X_3X_4 + \ldots + X_{2n-1}X_{2n}]}{X_1^2 + X_2^2 + \ldots + X_{2n}^2} \xrightarrow{L} \frac{Z}{2} \sim N(0, 1/4)$$

$$[\because Z \sim N(0, 1) \;\Rightarrow\; CZ \sim N(0, C^2)]$$

EXERCISE 8(g)

1. State and prove the central limit theorem for the sum of n independently and identically distributed random variables with positive finite variance under conditions to be stated.

2. State Lindeberg's sufficient conditions for the central limit theorem to hold for a sequence $\{X_k\}$ of independent random variables. Show that every uniformly bounded sequence $\{X_k\}$ of mutually independent random variables obeys the central limit theorem.

Comment on the case when the random variables do not possess expectations.

3. A distribution with unknown mean μ has variance equal to 1·5. Use central limit theorem to find how large a sample should be taken from the

distribution in order that the probability will be at least 0·95 that the sample mean will be within 0·5 of the population mean.

4. The life time of a certain brand of an electric bulb may be considered a random variable with mean 1,200 hours and standard deviation 250 hours. Find the probability, using central limit theorem, that the average life-time of 60 bulbs exceeds 1,400 hours.

5. State the Liapounoff form of central limit theorem.

Decide whether the central limit theorem holds for the sequence of independent random variables X_r with distribution defined as follows:

$$P(X_r = 1) = p_r \text{ and } P(X_r = 0) = 1 - p_r$$

6. Show that the central limit theorem applies if

(i) $P(X_k = \pm k^{\alpha}) = \frac{1}{2}$, (ii) $P(X_k = \pm\sqrt{2k-1}) = \frac{1}{2}$, and

(iii) $P(X_k = 0) = 1 - k^{1-2\alpha}, P(X_k = \pm k^{\alpha}) = \frac{1}{2}k^{-2\alpha}$, where $\alpha < \frac{1}{2}$

7. If $X_1, X_2, X_3 \ldots$ is a sequence of independent random variables having the uniform densities

$$f_i(x_i) = \begin{cases} 1/(2 - i^{-1}), & 0 < x_i < 2 - i^{-1} \\ 0 & \text{elsewhere}, \end{cases}$$

show that the central limit theorem holds.

8. Let X_n be the sample mean of a random sample of size n from Rectangular distribution on [0, 1]. Let

$$U_n = \sqrt{n}\ (\bar{X}_n - \tfrac{1}{2}).$$

Show that $$F(u) = \lim_{n \to \infty} P(U_n < u)$$

exists and determine it.

Ans. $\Phi(\sqrt{12}\, u)$, where $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-x^2/2}\, dx$

9. Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed non-negative random variables such that $E(\log X_1)^2$ is finite. $Z_n = (X_1 X_2 \ldots X_n)^{1/n}$. Show that the positive constant c can be so chosen that the random variable $(cZ_n)^{\sqrt{n}}$ has a non-degenerate limit distribution function $F(.)$ and determine $F(.)$.

Ans. $c = e^{-\mu}, F(x) = \Phi(\log x/\sigma), \mu = E(\log X_1)$ and $\sigma^2 = V(\log X_1)$.

10. $\{X_n\}$ is a sequence of *i.i.d.* random variables. If n is a perfect square, then X_n is a Cauchy variate with density $\frac{1}{\pi} \cdot \frac{1}{1+x^2}, -\infty < x < \infty$.

Otherwise X_n has a distribution function $F(x)$ with mean zero and finite variance σ^2. Discuss the asymptotic distribution of $(X_1 + X_2 + \ldots + X_n)/\sqrt{n}$.

11. Let $\{X_k\}$, $k \geq 1$ be a sequence of *i.i.d.* variates with

$$f(x) = \frac{1}{2} e^{-|x|}, -\infty < x < \infty.$$

Find the constants a_n and b_n such that

$$\left\{ |X_1| + |X_2| + \ldots + |X_n| - a_n \right\} / b_n \xrightarrow{d} N(0, 1)$$

[Indian Civil Services, 1982]

12. Using C.L.T, show that

$$\lim_{n \to \infty} \left[e^{-n} \sum_{k=0}^{n} \frac{n^k}{k!} \right] = \frac{1}{2} = \lim_{n \to \infty} \int_0^n \frac{e^{-t} \cdot t^{n-1}}{(n-1)!} dt$$

(Indian Civil Services, 1984)

13. Let $\{ X_n, n = 1, 2, \ldots \}$ be a sequence of independent Bernoulli variates such that: $P(X_n = 1) = p_n = 1 - P(X_n = 0),\ n = 1, 2, \ldots, (q_n = 1 - p_n)$.

Show that if $\Sigma p_n q_n = \infty$, $(n = 1, 2, \ldots, \infty)$, then the CLT holds for the sequence $\{ X_n \}$. What happens if $\Sigma p_n q_n < \infty$. (Indian Civil Services, 1988)

14. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed r.v.'s with $E(X_i) = \mu$; $\text{Var}(X_i) = \sigma^2$; $(0 < \sigma^2 < \infty)$; $i = 1, 2, \ldots, n$ and $E(X_i - \mu)^4 = \sigma^4 + 1$.

(*a*) State weak law of large numbers.

(*b*) If $P\left[\frac{1}{n} \left(X_1^2 + X_2^2 + \ldots + X_n^2 \right) - c \right] \to 0$ as $n \to \infty$, find c.

Hint. By Khinchines theorem $c = E X_i^2 = \sigma^2 + \mu^2$ (finite).

(*c*) State the Lnidberg–Levy Central Limit theorem.

(*d*) Find $\lim_{n \to \infty} P\left[\sigma^2 - \frac{1}{\sqrt{n}} \le \frac{(X_1 - \mu)^2 + \ldots + (X_n - \mu)^2}{n} \le \sigma^2 + \frac{1}{\sqrt{n}} \right]$

[Delhi Univ. B.A. (Stat. Hons.), Spl. Course 1986]

Hint.
$$p_n = P\left[-\frac{1}{\sqrt{n}} \le \frac{\Sigma (X_i - \mu)^2}{n} - \sigma^2 \le \frac{1}{\sqrt{n}} \right]$$

$$= P\left[-\sqrt{n} \le \Sigma (X_i - \mu)^2 - n\sigma^2 \le \sqrt{n} \right]$$

$$= P\left[-1 \le \sum_{i=1}^{n} \left[(X_i - \mu)^2 - \sigma^2 \right] / \sqrt{n} \le 1 \right]$$

$$= P\left[-1 \le S_n / \sqrt{n} \le 1 \right] \qquad \ldots(*)$$

where
$$S_n = \sum_{i=1}^{n} \left[(X_i - \mu)^2 - \sigma^2 \right] = \sum_{i=1}^{n} U_i$$

where $U_i = (X_i - \mu)^2 - \sigma^2,\ i = 1, 2, \ldots, n$; are *i.i.d. r.v.'s*.

$\Rightarrow$
$$E(U_i) = E(X_i - \mu)^2 - \sigma^2 = \sigma^2 - \sigma^2 = 0$$

$$\text{Var}\, U_i = \text{Var}\left[(X_i - \mu)^2 - \sigma^2 \right] = \text{Var}(X_i - \mu)^2$$

$$= E(X_i - \mu)^4 - [E(X_i - \mu)^2]^2 = \sigma^4 + 1 - \sigma^4 = 1$$

$$[\because E(X_i - \mu)^4 = \sigma^4 + 1 \text{ (Given)}]$$

$$\therefore \quad E(S_n) = \sum_{i=1}^{n} E(U_i) = 0 \ ; \ \text{Var}(S_n) = \text{Var}\left(\sum_{i=1}^{n} U_i\right) = \sum_{i=1}^{n} \text{Var}(U_i) = n$$

$$(\because U_i\text{'s are } i.i.d.)$$

Hence by C.L.T. $\dfrac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \dfrac{S_n}{\sqrt{n}} \xrightarrow{L} N(0, 1)$ as $n \to \infty$ ...(**)

From (*) and (**), we get

$$\lim_{n \to \infty} p_n = \lim_{n \to \infty} \left[-1 \le S_n/\sqrt{n} \le 1\right] = P(-1 \le Z \le 1), \text{ where } Z \sim N(0, 1)$$

$$= 2 \times P(0 \le Z \le 1) = 2 \times 0{\cdot}3413 = 0{\cdot}6826$$

15. Let $X_1, X_2, ..., X_n$ be i.i.d. $N(0, 1)$ variates. Show that the limiting distribution of

$$\sqrt{n}\,(X_1 + X_2 + \ldots + X_n)/(X_1^2 + X_2^2 + \ldots + X_n^2) \sim N(0, 1) \text{ as } n \to \infty.$$

Hint. Use Cramer's Theorem.

16. Let $X_1, X_2, ..., X_{2n}$ be i.i.d. $N(0, 1)$ variates. Find the limiting distribution of $Z_n = U_n/V_n$ where

$$U_n = \left(\frac{X_1}{X_2} + \frac{X_3}{X_4} + \ldots + \frac{X_{2n-1}}{X_{2n}}\right), \quad V_n = X_1^2 + X_2^2 + \ldots + X_n^2.$$

Hint. X_i are $i.i.d.$ $N(0, 1)$; $i = 1, 2, .., 2n$; $E(X_i^2) = \text{Var}\, X_i = 1$

$\Rightarrow (X_{2i-1}/X_{2i})$ are $i.i.d.$ standard Cauchy variates; $i = 1, 2, .., n$

$$\Rightarrow \lim_{n \to \infty} \frac{U_n}{n} \xrightarrow{L} \text{Standard Cauchy Variate} = C(0, 1),$$

(Being the mean of i.i.d. standard Cauchy variates)

$$\lim_{n \to \infty} \frac{V_n}{n} = \frac{X_1^2 + X_2^2 + \ldots + X_n^2}{n} \xrightarrow{P} E X_i^2 = 1,$$

(By Khinchine's WLLN)

$$\therefore \qquad \frac{U_n}{V_n} = \left(\frac{U_n}{n}\right) \bigg/ \left(\frac{V_n}{n}\right) \xrightarrow{L} C(0, 1)$$

(By Cramer's Theorem)

17. Let $\{X_n\}$ be a sequence of i.i.d. r.v.'s with mean α and variance σ^2 and let $\{Y_n\}$ be another sequence of i.i.d. r.v.'s with mean $\beta\ (\neq 0)$ and variance σ_1^2. Find the limiting distribution of :

$$Z_n = \sqrt{n}\ (\bar{X}_n - \alpha)/\bar{Y}_n \text{ where } \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \ \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Hint. $U_n = \dfrac{\bar{X}_n - E(\bar{X}_n)}{\sigma/\sqrt{n}} \xrightarrow{L} Z \sim N(0, 1)$ (By CLT)

$V_n = \bar{Y}_n \xrightarrow{P} E(Y_n) = \beta$ *(By WLLN)*

By Cramer's Theorem:

$$\lim_{n \to \infty} \frac{U_n}{V_n} = \frac{\sqrt{n}\,(\bar{X}_n - \alpha)}{\sigma \bar{Y}_n} \xrightarrow{L} \frac{Z}{\beta}, \text{ where } Z \sim N(0, 1)$$

$$\Rightarrow \quad \lim_{n \to \infty} \frac{\sqrt{n}\,(\bar{X}_n - \alpha)}{\bar{Y}_n} \xrightarrow{L} \frac{\sigma}{\beta} Z \sim N\left(0, \frac{\sigma^2}{\beta^2}\right).$$

18. Let n numbers $X_1, X_2, \ldots, X_n$ in decimal form, be each approximated by the closest integer. If X_i is the ith true number and Y_i is the nearest integer, then $U_i = X_i - Y_i$, is the error made by the rounding process. Suppose that $U_1, U_2, \ldots, U_n$ are independent and each is uniform on $(-0{\cdot}5, 0{\cdot}5)$.

(*i*) What is the probability that the true sum is within 'a' units of the approximated sum?

(*ii*) If $n = 300$ terms are added, find 'a' so that we are 95% sure that the approximation is within 'a' units of the true sum.

Hint. Reqd. Prob. 'p' $= P\left[\left|\sum_{i=1}^{n} (X_i - Y_i)\right| \le a\right] = P\left[-a \le \sum_{i=1}^{n} U_i \le a\right]$

Now use Lindeberg Levy C.L.T. for i.i.d. r.v.'s

$U_i \sim U[-0{\cdot}5, 0{\cdot}5]$ with $E(U_i) = 0$, Var $(U_i) = 1/12$

Ans. (*i*) $p = 2\,\Phi(a\sqrt{12/n}) - 1$;

(*ii*) $p = 0{\cdot}95 \Rightarrow \Phi\left(a\sqrt{12/300}\right) = 0{\cdot}975 \Rightarrow \frac{a}{5} = 1{\cdot}96 \Rightarrow a = 9{\cdot}8.$

8·11. Compound Distributions. Consider a random variable X whose distribution depends on a single parameter θ which instead of being regarded as a fixed constant, is also a random variable following a particular distribution. In this case, we say that the random variable X has a *compound or composed distribution.*

8·11·1. Compound Binomial Distribution. Let us suppose that $X_1, X_2, X_3, \ldots$ are identically and independently distributed Bernoulli variates with

$$P(X_i = 1) = p \text{ and } P(X_i = 0) = q = 1 - p$$

For a fixed n, the random variable $X = X_1 + X_2 + \ldots + X_n$ is a Binomial variate with parameters n and p and probability function:

$$P(X = r) = \binom{n}{r} p^r q^{n-r}\,;\; r = 0, 1, 2, \ldots, n$$

which gives the probability of r successes in n independent trials with constant probability 'p' of success for each trial.

Now suppose that n, instead of being regarded as a fixed constant, is also a random variable following Poisson law with parameter λ. Then

$$P(n = k) = \frac{e^{-\lambda} \lambda^k}{k!}\,;\; k = 0, 1, 2, \ldots$$

In such a case X is said to have compound binomial distribution. The joint probability function of X and n is given by

$$P(X = r \cap n = k) = P(n = k)\,P(X = r \mid n = k)$$

$$= \frac{e^{-\lambda}\lambda^k}{k!}\binom{k}{r}p^r q^{k-r},$$

since $P(X=r \mid n=k)$ is the probability of r successes in k trials. Obviously, $r \le k \Rightarrow k \ge r$.

The marginal distribution of X is given by

$$P(X=r) = \sum_{k=r}^{\infty} P(X=r \cap n=k)$$

$$= e^{-\lambda} p^r \sum_{k=r}^{\infty} \binom{k}{r} \frac{\lambda^k q^{k-r}}{k!} = \frac{e^{-\lambda}(\lambda p)^r}{r!} \sum_{k=r}^{\infty} \frac{(\lambda q)^{k-r}}{(k-r)!}$$

$$= \frac{e^{-\lambda}(\lambda p)^r}{r!} \sum_{j=0}^{\infty} \frac{(\lambda q)^j}{j!}, \quad (j = k-r)$$

$$= \frac{e^{-\lambda}(\lambda p)^r}{r!} \cdot e^{\lambda q} = \frac{e^{-\lambda p}(\lambda p)^r}{r!},$$

which is the probability function of a Poisson variate with parameter λp.

Hence $E(X) = \lambda p$ and $\text{Var}(X) = \lambda p$

We give below some of the practical situations where we would come across compound Binomial distribution :

1. Suppose that the probability of an insect laying n eggs is given by the Poisson distribution $e^{-\lambda}\lambda^n/n!$ and the probability of an egg developing is p. Assuming natural independence of eggs, the probability of a total of k survivors is given by the Poisson distribution with parameter λp.

2. The probability that a radioactive substance gives off n Beta particles in a unit of time is $P(\lambda)$, $(n = 0, 1, 2, ...)$. The probability that a given particle will strike a counter and be registered is p. Then the probability of registering n Beta particles in a unit of time is also $P(\lambda p)$.

3. If the probability of number of hits by lightning during any time interval t is $P(\lambda t)$ and if the probability of its hitting and damaging an individual is p then (assuming stochastic independence) the total damage during time 't' is $P(\lambda t p)$.

8·11·2. Compound Poisson Distribution. Let X be a $P(\lambda)$ so that

$$P(X=r) = \frac{e^{-\lambda}\lambda^r}{r!}\,; \; r = 0, 1, 2, \ldots$$

where λ itself is a continuous random variable with generalised gamma density

$$g(\lambda) = \begin{cases} \dfrac{a^{\nu}}{\Gamma(\nu)} e^{-a\lambda}\lambda^{\nu-1}\,; & \lambda > 0, a > 0, \nu > 0 \\ 0, & \lambda \le 0 \end{cases}$$

Let us consider the two dimensional random vector (X, λ) in which one variable is discrete and the other is continuous. For a constant $h > 0$ and $\lambda_1 > 0$, the joint density of X and λ is given by

$$P(X=r\cap\lambda_1\le\lambda\le\lambda_1+h)=P(\lambda_1\le\lambda\le\lambda_1+h)\,P(X=r\mid\lambda_1\le\lambda\le\lambda_1+h)$$

Dividing both sides by h and proceeding to the limits as $h\to 0$, we get

$$\lim_{h\to 0}\frac{P(X=r\cap\lambda_1\le\lambda\le\lambda_1+h)}{h}=\lim_{h\to 0}P(X=r\mid\lambda_1\le\lambda\le\lambda_1+h)\times\lim_{h\to 0}\frac{P(\lambda_1\le\lambda\le\lambda_1+h)}{h}$$

But

$$\lim_{h\to 0}\frac{P(\lambda_1\le\lambda\le\lambda_1+h)}{h}=\lim_{h\to 0}\frac{G(\lambda_1+h)-G(\lambda_1)}{h}=G'(\lambda_1)=g(\lambda_1)$$

where $G(\,.\,)$ is the distribution function and $g(\,.\,)$ is the p.d.f. of λ.

$$\therefore\ \lim_{h\to 0}\frac{P(X=r\cap\lambda_1\le\lambda\le\lambda_1+h)}{h}=\frac{e^{-\lambda_1}\lambda_1^r}{r\,!}\cdot\frac{a^v}{\Gamma(v)}\lambda_1^{v-1}e^{-a\lambda_1}$$

Integrating w.r.to λ_1 over 0 to ∞ and using gamma integral, the marginal probability function of X is given by

$$P(X=r)=\frac{a^v}{\Gamma(v)\,r\,!}\int_0^\infty e^{-(1+a)\lambda}\lambda^{r+v-1}\,d\lambda$$

$$=\frac{a^v}{\Gamma(v)\,r\,!}\cdot\frac{\Gamma(r+v)}{(1+a)^{r+v}}$$

$$=\left(\frac{a}{1+a}\right)^v\frac{v(v+1)(v+2)\ldots(v+r-1)}{(1+a)^r\,r\,!}$$

$$=\left(\frac{a}{1+a}\right)^v(-1)^r\binom{-v}{r}\left(\frac{1}{1+a}\right)^r$$

$$=\binom{-v}{r}p^v(-q)^r\ ;\ r=0,1,2,\ldots$$

where $p=a/(1+a)$, $q=1-p=1/(1+a)$

Thus the marginal distribution of X is a negative binomial with parameters (v,p).

EXERCISE 8(h)

1. (*a*) What do you mean by a compound distribution? Obtain the probability function of compound Poisson distribution and identify it.

(*b*) What is the Compound Binomial distribution? Obtain its probability function and identify the distribution.

2. If X is a random variable with p.d.f.,

$$f(x)=\frac{1}{\Gamma(n+1)}e^{-x}x^n,\ x\ge 0$$

where n is a positive integer, and the discrete random variable Y has a Poisson distribution with parameter λ, show that $P(X\ge\lambda)=P(Y\le n)$.

Hint. $$P(X \geq \lambda) = 1 - \frac{1}{\Gamma(n+1)} \int_0^{\lambda} e^{-x} x^n \, dx$$

Integrating by parts successively, we get the result.

3. If X has a Poisson distribution:

$$P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}; \; r = 0, 1, 2, \ldots$$

where the parameter λ is a random variable of the continuous type with the density function:

$$f(\lambda) = \frac{a^{\nu}}{\Gamma(\nu)} \cdot e^{-a\lambda} \lambda^{(\nu-1)}; \; \lambda \geq 0, a > 0, \nu > 0,$$

derive the distribution of X.

Show that the characteristic function of X is given by

$$\Phi_X(t) = E(e^{itX}) = q^{\nu}(1 - pe^{it})^{-\nu}, \text{ where } p = 1/(1+a), q = 1-p$$

[South Gujarat Univ. M.Sc. 1991]

4. The conditional distribution of a continuous random variable X for a discrete random variable Y, assuming a value n is

$$dF = n(1-x)^{n-1} \, dx, \; 0 \leq x \leq 1$$

The distribution of Y is: $P(Y = n) = (\frac{1}{2})^n$; $n = 1, 2, 3, \ldots$

Find the marginal distribution of X.

5. Given $f(x \mid y) = \dfrac{e^{-y} \cdot y^x}{x!}$ and $h(y) = e^{-y}$, where X is discrete, *i.e.*, $x = 0, 1, 2, \ldots$ and Y is continuous, $y \geq 0$; show that the marginal distribution of X is geometric, *i.e.*, $g(x) = (\frac{1}{2})^{x+1}$.

6. The conditional probability that the random variable X should lie within the range dx for a given σ is given by

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}(x-\mu)^2/\sigma^2\right\} dx, \; -\infty < x < \infty$$

while the probability of σ itself lying within the range $d\sigma$ is

$$\frac{1}{\sigma_0^2} \exp\left\{-\tfrac{1}{2}\sigma^2/\sigma_0^2\right\} \sigma \, d\sigma, \; 0 < \sigma < \infty$$

where σ_0 is a constant. Show that the unconditional (*i.e.*, marginal) distribution of X has the following probability function:

$$\frac{1}{2\sigma_0} \exp\left\{-(1/\sigma_0)\,|x - \mu|\right\}, \; -\infty < x < \infty$$

7. Let $X \sim U[0, 1]$ and $Y \mid (X = x) \sim B(n, x)$ i.e.,

$$P(Y = y \mid X = x) = \binom{n}{y} x^y (1-x)^{n-y}, \; y = 0, 1, 2, \ldots, n$$

Find the distribution of Y. Also find $E(Y)$.

Ans. $P(Y = y) = 1/(n+1), \; y = 0, 1, \ldots, n \Rightarrow Y \sim U[0, 1, 2, \ldots, n]; \; E(Y) = n/2$

8·12. Pearson's Distributions. Given a set of observations from a population, the first question that arises in our mind is about the nature of the parent population. A vague idea is provided by the frequency polygon (or frequency curve) but the information is totally inadequate and unreliable, because the sample observations may not cover the entire range of the parent distribution. Moreover, an unusually high frequency in one class, arising out of sheer chance, may completely distort the shape of the frequency curve.

Consequently, to determine the frequency curve, we resort to the technique of curve-fitting to the given data. The failure of the normal distribution to fit many distributions which are observed in practice for continuous variables necessitated the development of generalised system of frequency curves. Since a trial and error approach is clearly undesirable, an elastic system of frequency curves must be evolved, which should incorporate, if not all, at least the most common of the distributions. *Pearsonian system of frequency curves* is one of the most important approaches in this direction, in which we decide about the shape of the curve on the basis of a '*criterion* κ' calculated from the sample observations.

Karl Pearson's first memoir dealing with generalised frequency curves appeared in 1895. In this paper and the subsequent two papers published in 1908 and 1916, Karl Pearson developed a set of frequency curves which could be obtained by assigning values to the parameters in a certain first order differential equation.

Genesis of Pearson's Frequency Curves. Experience tells us that most of the frequency distributions possess the following obvious and common characteristic:

"They rise from a low frequency to a maximum frequency and then again fall to the low frequency as the variable X increases. This suggests a unimodal frequency curve $y = f(x)$ with high contact at the extremities of the range, *i.e.*, $\dfrac{dy}{dx} = 0$ when $y = 0$. Accordingly, Karl Pearson proposed the following differential equation for the frequency curve $y = f(x)$,

$$\frac{dy}{dx} = \frac{y(x-a)}{F(x)} \qquad \ldots(8\cdot32)$$

where $F(x)$ is an arbitrary function of x not vanishing at $x = a$, the mode of the distribution. Expanding $F(x)$ by Maclaurin's theorem, we get $F(x) = b_0 + b_1 x + b_2 x^2 + \ldots$ and retaining only the first three terms we get the differential equation of the Pearsonian system of frequency curves as

$$\frac{dy}{dx} = \frac{y(x-a)}{b_0 + b_1 x + b_2 x^2} \Rightarrow \frac{df(x)}{dx} = f'(x) = \frac{(x-a)f(x)}{b_0 + b_1 x + b_2 x^2} \qquad \ldots(8\cdot32\,a)$$

where a, b_0, b_1 and b_2 are the constants to be calculated from the given data.

Remark. Equation (8·32 *a*) can also be obtained as a limiting case of Hyper-geometric distribution (*c.f. Advanced statistics Vol. II by Kendall*).

8·12·1. Determination of the Constants of the Equation in Terms of Moments. Multiplying both sides of (8·32 *a*) by x^n for integral $n > 0$ and integrating over the entire range of the variable X say (α, β), we get

$$\int_\alpha^\beta x^n (b_0 + b_1 x + b_2 x^2) f'(x)\, dx = \int_\alpha^\beta x^n (x - a) f(x)\, dx$$

$$\Rightarrow \left| x^n (b_0 + b_1 x + b_2 x^2) f(x) \right|_\alpha^\beta - \int_\alpha^\beta \left[n b_0 x^{n-1} + (n+1) b_1 x^n + (n+2) b_2 x^{n+1} \right] f(x)\, dx$$
$$= \int_\alpha^\beta x^{n+1} f(x)\, dx - a \int_\alpha^\beta x^n f(x)\, dx$$

Assuming high order contact at the extremities so that

$$\left| x^r f(x) \right|_\alpha^\beta = 0 \text{ i.e., } x^r f(x) \to 0 \text{ as } x \to \alpha \text{ or } \beta, \text{ we get} \qquad \ldots(*)$$

$$-\left[n b_0 \mu_{n-1} + (n+1) b_1 \mu_n + (n+2) b_2 \mu_{n+1} \right] = \mu_{n+1} - a \mu_n$$

(assuming that X is measured from mean and this we can do without any loss of generality). Thus the recurrence relation between the moments becomes

$$n b_0 \mu_{n-1} + \left[(n+1) b_1 - a \right] \mu_n + \left[(n+2) b_2 + 1 \right] \mu_{n+1} = 0, \qquad \ldots(**)$$
$$n = 1, 2, 3, \ldots$$

Integrating (8·32 *a*) w.r.to x within the limits (α, β) and using (*), we get

$$(b_1 - a) + (2b_2 + 1) = 0 \qquad \ldots(***)$$

Putting $n = 1, 2,$ and 3 in (**) and solving these equations and (***) with the help of determinants and using $\mu_0 = 1, \mu_1 = 0$, we get

$$b_0 = -\frac{\mu_2 (4 \mu_2 \mu_4 - 3 \mu_3^2)}{2(5 \mu_2 \mu_4 - 9 \mu_2^3 - 6 \mu_3^2)} = -\frac{\sigma^2 (4 \beta_2 - 3 \beta_1)}{2 (5 \beta_2 - 6 \beta_1 - 9)}$$

$$a = b_1 = -\frac{\mu_3 (\mu_4 + 3 \mu_2^2)}{2 (5 \mu_2 \mu_4 - 9 \mu_2^3 - 6 \mu_3^2)} = -\frac{\sigma \sqrt{\beta_1} (\beta_2 + 3)}{2 (5 \beta_2 - 6 \beta_1 - 9)} \qquad \ldots(8·33)$$

$$b_2 = -\frac{(2 \mu_2 \mu_4 - 3 \mu_2^3 - 6 \mu_3^2)}{2 (5 \mu_2 \mu_4 - 9 \mu_2^3 - 6 \mu_3^2)} = -\frac{(2 \beta_2 - 3 \beta_1 - 6)}{2 (5 \beta_2 - 6 \beta_1 - 9)}$$

where $\mu_2 = \sigma^2$, $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$.

Thus the Pearsons's system (8·32 *a*) is completely specified by the first four moments.

8·12·2. Pearson Measure of Skewness.

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{0 - a}{\sqrt{\mu_2}}$$

$$= \frac{\sqrt{\beta_1} (\beta_2 + 3)}{2 (5 \beta_2 - 6 \beta_1 - 9)} \qquad \ldots(8·34)$$

8·12·3. "Criterion κ". Equation (8·32 *a*) can be re-written as:

$$\frac{df}{f}=\frac{(x-a)\,dx}{b_0+b_1x+b_2x^2},\quad f=f(x)$$

Integrating we get

$$\log(f/c)=\int\frac{(x-a)}{b_0+b_1x+b_2x^2}\,dx=I\text{ (say)},$$

where c is the constant of integration.

$$\therefore\qquad f(x)=c\exp\,[I]$$

Thus f depends on I, which further depends on the roots of the equation

$$b_0+b_1x+b_2x^2=0 \qquad \ldots(**)$$

Now,

$$b_0+b_1x+b_2x^2=b_2\left[x^2+\frac{b_1}{b_2}x+\frac{b_0}{b_2}\right]$$

$$=b_2\left[x-\frac{-b_1+\sqrt{b_1^2-4b_0b_2}}{2b_2}\right]\left[x-\frac{-b_1-\sqrt{b_1^2-4b_0b_2}}{2b_2}\right]$$

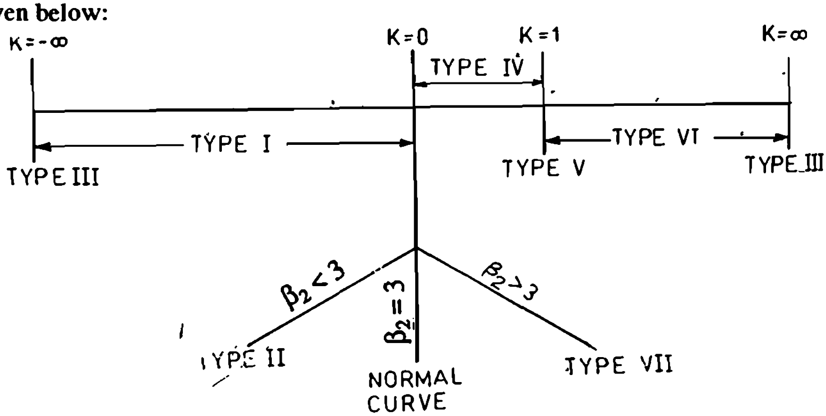
$$=b_2\left[x+\frac{b_1}{2b_2}-\frac{\sqrt{b_1^2-4b_0b_2}}{\sqrt{4b_2^2}}\right]\left[x+\frac{b_1}{2b_2}+\frac{\sqrt{b_1^2-4b_0b_2}}{\sqrt{4b_2^2}}\right]$$

$$=b_2\left[x+\frac{b_1}{2b_2}-\sqrt{\frac{b_0}{b_2}(\kappa-1)}\right]\left[x+\frac{b_1}{2b_2}+\sqrt{\frac{b_0}{b_2}(\kappa-1)}\right]$$

where $\kappa=b_1^2/(4b_0b_2)$, ...(8·35)

determines the criterion for obtaining the form of the frequency curve.

A brief description of various Pearsonian curves for different values of κ is given below:



8·12·4. Pearson's Main Type 1. *This curve is obtained when the roots of the quadratic equation are real and of opposite sign, i.e., when* $\kappa<0$.

Shifting the origin to the mode $x=a$, the equation becomes

$$\frac{df}{dx}=\frac{xf}{B_0+B_1x+B_2x^2}=\frac{xf}{B_2(x+\alpha)(x-\beta)}$$

where $B_0=b_0, B_1=b_1$ and $B_2=b_2$.

$$\Rightarrow \quad \frac{d}{dx}(\log f)=\frac{1}{B_2(\alpha+\beta)}\left[\frac{\alpha}{x+\alpha}+\frac{\beta}{x-\beta}\right]$$

Integrating both sides w.r.to. x, we get

$$\log f=\log(x+\alpha)^{\alpha/B_2(\alpha+\beta)}+\log(x-\beta)^{\beta/B_2(\alpha+\beta)}+\log C$$

$$\therefore \quad f=C(x+\alpha)^{\alpha/B_2(\alpha+\beta)}(x-\beta)^{\beta/B_2(\alpha+\beta)}$$

$$\Rightarrow \quad f=y_0\left(1+\frac{x}{\alpha}\right)^{\alpha/B_2(\alpha+\beta)}\left(1-\frac{x}{\beta}\right)^{\beta/B_2(\alpha+\beta)}, -\alpha\le x\le\beta \qquad ...(*)$$

Let $\alpha=a_1, \beta=a_2, m_1=\dfrac{\alpha}{B_2(\alpha+\beta)}$ and $m_2=\dfrac{\beta}{B_2(\alpha+\beta)}$ so that

$\dfrac{m_1}{a_1}=\dfrac{m_2}{a_2}=\dfrac{1}{B_2(a_1+a_2)}$, then (*) may be written as

$$f(x)=y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}, -a_1\le x\le a_2 \qquad ...(8{\cdot}36)$$

which is a standard form of Type 1.

Determination of y_0:

$$1=y_0\int_{-a_1}^{a_2}\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}dx$$

Put $x=(a_1+a_2)z-a_1$ so that $dx=(a_1+a_2)dz$

$$\Rightarrow 1=y_0\int_0^1\frac{(a_1+a_2)^{m_1}}{a_1^{m_1}}z^{m_1}\frac{(a_1+a_2)^{m_2}}{a_2^{m_2}}(1-z)^{m_2}(a_1+a_2)\,dz$$

$$\Rightarrow \quad 1=y_0\left[\frac{(a_1+a_2)^{m_1+m_2+1}}{a_1^{m_1}a_2^{m_2}}\right]B(m_1+1, m_2+1)$$

$$\therefore \quad y_0=\frac{a_1^{m_1}a_2^{m_2}}{(a_1+a_2)^{m_1+m_2+1}B(m_1+1, m_2+1)}$$

Remark. It may be noted that Beta distribution is a particular case of Type I distribution.

Determination of moments:

$$\mu_n'=y_0\int_{-a_1}^{a_2}(x+a_1)^n\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}dx$$

$$=y_0\frac{(a_1+a_2)^{m_1+m_2+n+1}}{a_1^{m_1}a_2^{m_2}}B(n+m_1+1, m_2+1),$$

where $x=(a_1+a_2)z-a_1$

$$= \frac{(a_1 + a_2)^n}{B(m_1 + 1, m_2 + 1)} \cdot B(n + m_1 + 1, m_2 + 1)$$

[On simplification]

8·12·5. Pearson's Type IV. *This curve is obtained when the roots are imaginary or when*

$$B_1^2 < 4 B_0 B_2, \quad i.e., \quad 0 < \kappa < 1$$

$$\frac{1}{f} \cdot \frac{df}{dx} = \frac{x}{B_0 + B_1 x + B_2 x^2} \qquad \text{[Origin at mode]}$$

$$\Rightarrow \quad \frac{d}{dx}(\log f) = \frac{x}{B_2 \left[\left(x + \frac{B_1}{2 B_2} \right)^2 + \left(\frac{B_0}{B_2} - \frac{B_1^2}{4 B_2^2} \right) \right]}$$

$$= \frac{(x + \gamma) - \gamma}{B_2 [(x + \gamma)^2 + \delta^2]} = \frac{2(x + \gamma)}{2 B_2 [(x + \gamma)^2 + \delta^2]} - \frac{2\gamma}{2 B_2 [(x + \gamma)^2 + \delta^2]}$$

$$\log f = \frac{1}{2 B_2} \log [(x + \gamma)^2 + \delta^2] - \frac{\gamma}{B_2} \cdot \frac{1}{\delta} \tan^{-1} \frac{x + \gamma}{\delta} + \log C$$

$$\therefore \quad f = C \left[(x + \gamma)^2 + \delta^2 \right]^{\frac{1}{2 B_2}} \exp \left\{ - \frac{\gamma}{B_2 \delta} \tan^{-1} \frac{x + \gamma}{\delta} \right\}$$

Put $\quad \dfrac{x + \gamma}{\delta} = \dfrac{x}{a}$ and $\dfrac{\gamma}{B_2 \delta} = \nu$

Hence $\quad f(x) = y_0 \left(1 + \dfrac{x^2}{a^2} \right)^{-m} e^{-\nu \tan^{-1}(x/a)}; \; -\infty < x < \infty, (m, \nu) > 0$ ...(8·37)

which is a standard form of Type IV with origin at $\left(-\dfrac{B_1}{2 B_2}, 0 \right)$. The curve is skew and has unltimited range in both the directions.

Determination of y_0 :

$$1 = y_0 \int_{-\infty}^{\infty} \left(1 + \frac{x^2}{a^2} \right)^{-m} e^{-\nu \tan^{-1}(x/a)} \, dx \qquad [\text{ Put } x = a \tan \theta]$$

$$= a y_0 \int_{-\pi/2}^{\pi/2} \cos^{2m-2} \theta \, e^{-\nu \theta} \, d\theta = a y_0 F(2m - 2, \nu)$$

$$\therefore \quad y_0 = \frac{1}{a F(2m - 2, \nu)}$$

Hence $\quad f(x) = \dfrac{1}{a F(2m - 2, \nu)} \left(1 + \dfrac{x^2}{a^2} \right)^{-m} e^{-\nu \tan^{-1}(x/a)}$

8·12·6. Pearson's Type VI. *The curve is obtained when the roots are real and are of the same sign. This is obtained when B_0, B_2 are of the same sign or, in other words, when $\kappa > 1$.*

Let $-\alpha_1$ and $-\alpha_2$ be the roots of the quadratic equation. Then

$$\frac{d}{dx}(\log f) = \frac{x}{B_0 + B_1 x + B_2 x^2} = \frac{x}{B_2 (x + \alpha_1)(x + \alpha_2)}$$

$$= \frac{\alpha_1}{B_2(\alpha_1 - \alpha_2)} \cdot \frac{1}{(x + \alpha_1)} - \frac{\alpha_2}{B_2(\alpha_1 - \alpha_2)} \cdot \frac{1}{(x + \alpha_2)}$$

$$\therefore \quad \log f = \frac{\alpha_1}{B_2(\alpha_1 - \alpha_2)} \log(x + \alpha_1) - \frac{\alpha_2}{B_2(\alpha_1 - \alpha_2)} \log(x + \alpha_2) + \log C$$

$$\Rightarrow \quad f = C(x + \alpha_1)^{\alpha_1/B_2(\alpha_1 - \alpha_2)} \cdot (x + \alpha_2)^{-\alpha_2/B_2(\alpha_1 - \alpha_2)}$$

Hence the probability density function is:

$$f(x) = y_0 \left(1 + \frac{x}{a_1}\right)^{m_1} \left(1 + \frac{x}{a_2}\right)^{-m_2} \qquad \ldots(8\cdot38)$$

where $\alpha_1 = a_1$, $\alpha_2 = a_2$ and $\dfrac{m_1}{a_1} = -\dfrac{m_2}{a_2}$; $a_1, a_2 > 0$

This equation can also be written (on shifting the origin to $-\alpha_2$ or $-a_2$) as

$$f(x) = y_0 (x - a)^{q_2} x^{-q_1}, \quad a \le x < \infty \qquad \ldots(8\cdot38\ a)$$

where $q_1 = \dfrac{\alpha_2}{B_2(\alpha_1 - \alpha_2)}$, $q_2 = \dfrac{\alpha_1}{B_2(\alpha_1 - \alpha_2)}$, $a = -\alpha_1$

Remark. The curve is bell shaped if $q_2 > 0$ and J-shaped if $q_2 < 0$.

Determination of y_0:

$$1 = y_0 \int_a^\infty (x - a)^{q_2} x^{-q_1} dx \qquad \left[\text{Put } x = \frac{a}{1 - z}\right]$$

$$= y_0 \int_0^1 \left(\frac{a}{1 - z}\right)^{-q_1} \left[\frac{a}{1 - z} - a\right]^{q_2} \frac{a}{(1 - z)^2} dz$$

$$= y_0 a^{q_2 - q_1 + 1} \int_0^1 z^{q_2} \frac{1}{(1 - z)^{q_2 - q_1 + 2}} dz$$

$$\therefore \quad y_0 = \frac{a^{q_1 - q_2 - 1}}{B(q_2 + 1, q_1 - q_2 - 1)}$$

Hence $f(x) = \dfrac{a^{q_1 - q_2 - 1}}{B(q_2 + 1, q_1 - q_2 - 1)} x^{-q_1} (x - a)^{q_2}, \ a \le x < \infty$

The above discussion covers almost the whole range of κ but in limiting cases we get simple cases. The following are more important of the transition curves when one of the main type changes into another.

8·12·7. Type III. *This is a transition type curve and is obtained when* $B_2 = 0, B_1 \neq 0$ *or* $\kappa \to \pm\infty$.

$$\frac{d}{dx}(\log f) = \frac{x}{B_0 + B_1 x}, \quad \text{(origin is at mode)}.$$

$$= \frac{B_1 x + B_0 - B_0}{B_1 (B_0 + B_1 x)} = \frac{1}{B_1} - \frac{B_0}{B_1 (B_0 + B_1 x)}$$

$\therefore$
$$\log f = \frac{x}{B_1} - \frac{B_0}{B_1^2} \log (B_0 + B_1 x) + \text{const.}$$

$$= \text{const. } e^{x/B_1} (B_0 + B_1 x)^{-B_0/B_1^2}$$

$$f(x) = y_0 \left(1 + \frac{x}{a} \right)^p e^{-px/a}; -a \leq x < \infty, \qquad \ldots(8{\cdot}39)$$

where
$$\frac{B_0}{B_1} = a \text{ and } \frac{B_0}{B_1^2} = p$$

This gives the Type III curve with origin at mode. The curve is usually bell shaped but becomes J-shaped when $\beta_1 > 4$.

Remark. The distribution can be transformed into the gamma form by using the transformation $y = \frac{p}{a}(x + a)$, when the curve reduces to

$$f(y) = \frac{1}{\Gamma(p + 1)} e^{-y} y^p, 0 \leq y < \infty$$

8·12·8. Type V. *This transition type is obtained when the roots are equal, i.e., when $B_1^2 = 4 B_0 B_2$ or $\kappa = 1$.*

$$\frac{d}{dx}(\log f) = \frac{x}{B_2 \left[\left(x + \frac{B_1}{2 B_2} \right)^2 \right]} = \frac{2 \left[x + \frac{B_1}{2 B_2} \right] - \frac{B_1}{B_2}}{2 B_2 \left[\left(x + \frac{B_1}{2 B_2} \right)^2 \right]}.$$

$\therefore$
$$\log f = \frac{1}{2 B_2} \log \left(x + \frac{B_1}{2 B_2} \right)^2 + \frac{B_1}{2 B_2^2} \frac{1}{\left[x + \frac{B_1}{2 B_2} \right]} + \text{const.}$$

$\Rightarrow$
$$f = \text{const} \left(x + \frac{B_1}{2 B_2} \right)^{\frac{1}{B_2}} \exp \left[\frac{B_1}{2 B_2^2} \left(x + \frac{B_1}{2 B_2} \right)^{-1} \right]$$

$\therefore$
$$f(x) = y_0 X^{-p} e^{-q/X}, \; 0 \leq X < \infty \qquad \ldots(8{\cdot}40)$$

where
$$X = \left(x + \frac{B_1}{2 B_2} \right), \frac{B_1}{2 B_2^2} = -q \text{ and } \frac{1}{B_2} = -p.$$

8·12·9. Type II. *This curve is obtained when $B_1 = 0$ and B_0, B_2 are of opposite sign, i.e., $\kappa = 0$. The equation to the curve is*

$$f(x) = y_0 \left[1 - \frac{x^2}{a^2} \right]^m, \; -a \leq x \leq a; \qquad \ldots(8{\cdot}41)$$

where $m = \frac{1}{2B_2} > 0, a^2 = -\frac{B_0}{B_2}$

with origin at mean (mode).

8·12·10. Type VII. *This curve is obtained when $B_1 = 0$ and B_0, B_2 are of the same sign, i.e.,* $\kappa = 0$ and $B_0 B_2 > 0$. The equation to the curve is

$$f(x) = y_0 \left[1 + \frac{x^2}{a^2}\right]^{-m}, -\infty < x < \infty \qquad ...(8·42)$$

where $a^2 = \frac{B_0}{B_2}$ and $m = -\frac{1}{2B_2}$

with origin being at the mean (mode). This curve is usually bell shaped, symmetrical and of unlimited range in both the directions.

8·12·11. Zero Type (Normal curve). When $B_1 = B_2 = 0$, (18·33) implies that $\beta_1 = 0$, and $\beta_2 = 3$ and we have

$$\frac{d}{dx}(\log f) = \frac{x}{B_0} \quad \Rightarrow \quad \log f = \frac{x^2}{2B_0} + \log C,$$

where C is the constant of integration.

$$\therefore \quad f = C \exp(x^2/2B_0) = C \exp(-x^2/2\sigma^2), -\infty < x < \infty \qquad ...(8·43)$$

where $B_0 = -\sigma^2$ and the origin is at mean. This is the normal distribution with mean zero and variance σ^2.

8·12·12. Type VIII. When $B_0 = 0, B_1 > 0$,

$$f(x) = \frac{1-m}{a}\left(1 + \frac{x}{a}\right)^m, -a \le x \le 0 \qquad ...(8·44)$$

Type IX. When $B_0 = 0, B_1 < 0$ and $\kappa < 0$

$$f(x) = \frac{1+m}{a}\left(1 + \frac{x}{a}\right)^m, -a \le x \le 0 \qquad ...(8·45)$$

Type X. When $B_0 = 0$ and $B_2 = 0$,

$$f(x) = \frac{1}{\sigma} e^{-x/\sigma}, 0 \le x < \infty, \sigma > 0 \qquad ...(8·46)$$

This is the p.d.f. of simple exponential distribution with parameter $\sigma > 0$.

Type XI. When $B_0 = B_1 = 0$, and $\kappa > 1$

$$f(x) = b^{m-1}(m-1)x^{m-1}, b \le x < \infty$$

Type XII. When $5\beta_2 - 6\beta_1 - 9 = 0, \kappa < 0$

$$f(x) = \left(\frac{a_1}{a_2}\right)^m \frac{1}{(a_1 + a_2)B(1+m, 1-m)} \cdot \frac{\left(1 + \frac{x}{a_1}\right)^m}{\left(1 - \frac{x}{a_2}\right)^m}, a_1 \le x \le a_2$$

Example 8·47. *Show that for a Pearson distribution :*

$$\frac{df}{f}=\frac{(a+x)\,dx}{b_0+b_1x+b_2x^2},$$

the characteristic function φ *obeys the relation :*

$$b_2\,\theta\frac{d^2\varphi}{d\theta^2}+(1+2b_2+b_1\,\theta)\frac{d\varphi}{d\theta}+(a+b_1+b_0\,\theta)\,\varphi=0,\ \textit{where}\ \theta=it.$$

Deduce the recurrence relation for moments.

Show also that the cumulant generating function ψ *obeys the relation :*

$$b_2\,\theta\left\{\frac{d^2\psi}{d\theta^2}+\left(\frac{d\psi}{d\theta}\right)^2\right\}+(1+2b_2+b_1\,\theta)\frac{d\psi}{d\theta}+(a+b_1+b_0\,\theta)=0,$$

Hence show that the cumulants obey the recurrence relation :

$$\left\{1+(r+2)\,b_2\right\}\kappa_{r+1}+rb_1\,\kappa_r+rb_2\left\{\binom{r-1}{1}\kappa_2\,\kappa_{r-1}+\binom{r-1}{2}\kappa_3\,\kappa_{r-2}\right.$$
$$\left.+\ldots+\binom{r-1}{j}\kappa_{j+1}\,\kappa_{r-j}+\ldots+\binom{r-1}{r-2}\kappa_{r-1}\,\kappa_2\right\}=0$$

Solution. $(b_0+b_1x+b_2x^2)\dfrac{df}{dx}=(a+x)f$

$$\Rightarrow\quad e^{\theta x}(b_0+b_1x+b_2x^2)\frac{df}{dx}=e^{\theta x}(a+x)f$$

Integrating w.r.to. x, for the total range of x, assuming that integrals vanish at either limit, we get

$$\int_{-\infty}^{\infty}e^{\theta x}(b_0+b_1x+b_2x^2)\frac{df}{dx}\cdot dx=\int_{-\infty}^{\infty}e^{\theta x}(a+x)f\,dx$$

$$\Rightarrow\quad \Big[e^{\theta x}(b_0+b_1x+b_2x^2)f\Big]_{-\infty}^{\infty}-\int_{-\infty}^{\infty}\Big\{\theta\,e^{\theta x}(b_0+b_1x+b_2x^2)$$
$$+e^{\theta x}(b_1+2\,b_2x)\Big\}f\,dx=\int_{-\infty}^{\infty}e^{\theta x}(a+x)f\,dx$$

$$\Rightarrow\quad 0-\theta\left[b_0\,\varphi+b_1\frac{d\varphi}{d\theta}+b_2\frac{d^2\varphi}{d\theta^2}\right]-\left[b_1\,\varphi+2b_2\frac{d\varphi}{d\theta}\right]=\left(a\,\varphi+\frac{d\varphi}{d\theta}\right)^*$$

---

* $\because\ \varphi=E(e^{itx})=\int_{-\infty}^{\infty}e^{itx}f\,dx=\int_{-\infty}^{\infty}e^{\theta x}f\,dx,$ $[\because it=\theta]$

$\dfrac{d\varphi}{d\theta}=\int_{-\infty}^{\infty}xe^{\theta x}f\,dx$ and $\dfrac{d^2\varphi}{d\theta^2}=\int_{-\infty}^{\infty}x^2e^{\theta x}f\,dx$

assuming differentiation is valid under integral sign.

$$\Rightarrow \quad b_2 \theta \frac{d^2 \varphi}{d \theta^2} + (1 + 2b_2 + b_1 \theta) \frac{d \varphi}{d \theta} + (a + b_1 + b_0 \theta) \varphi = 0 \text{ (On simplification)} \qquad \ldots(1)$$

Differentiating n times w.r.t. θ, using Leibnitz Theorem, we get

$$b_2 \left[\theta \frac{d^{n+2} \varphi}{d \theta^{n+2}} + n \cdot \frac{d^{n+1} \varphi}{d \theta^{n+1}} \cdot 1 \right] + \left[\left\{ \frac{d^{n+1} \varphi}{d\theta^{n+1}} (1 + 2b_2 + b_1 \theta) + \frac{d^n \varphi}{d \theta^n} \cdot nb_1 \right] \right.$$
$$+ \left[\left\{ \frac{d^n \varphi}{d \theta^n} (a + b_1 + b_0 \theta) + \left\{ \frac{d^{n-1} \varphi}{d \theta^{n-1}} \cdot nb_0 \right\} \right] = 0 \qquad \ldots(2)$$

Putting $\theta = 0$ and using the relation $\left[\dfrac{d^n \varphi}{d \theta^n} \right]_{\theta = 0} = \mu_n'$, we get

$$nb_2 \mu'_{n+1} + (2b_2 + 1) \mu'_{n+1} + nb_1 \mu_n' + (b_1 + a) \mu_n' + nb_0 \mu'_{n-1} = 0$$

Shifting the origin to the mean, we get

$$[(n + 2) b_2 + 1] \mu_{n+1} + [(n + 1) b_1 + a] \mu_n + n b_0 \mu_{n-1} = 0 \qquad \ldots(3)$$

Now $\varphi = e^{\psi}$, $\dfrac{d \varphi}{d \theta} = e^{\psi} \dfrac{d \psi}{d \theta}$ and $\dfrac{d^2 \varphi}{d \theta^2} = e^{\psi} \left[\dfrac{d^2 \psi}{d \theta^2} + \left(\dfrac{d \psi}{d \theta} \right)^2 \right]$

$(\because \psi = \log \varphi)$

Substituting these values in (1) and on simplification, we get

$$b_2 \theta \left[\frac{d^2 \psi}{d \theta^2} + \left(\frac{d \psi}{d \theta} \right)^2 \right] + (1 + 2 b_2 + b_1 \theta) \frac{d \psi}{d \theta} + (a + b_1 + b_0 \theta) = 0 \qquad \ldots(4)$$

Differentiating (4) r times w.r.to. θ using Leibnitz Theorem, we get

$$b_2 \theta \left[\frac{d^{r+2} \psi}{d \theta^{r+2}} + \frac{d^r}{d \theta^r} \left(\frac{d \psi}{d \theta} \right)^2 \right] + \binom{r}{1} b_2 \left[\frac{d^{r+1} \psi}{d \theta^{r+1}} + \frac{d^{r-1}}{d \theta^{r-1}} \left(\frac{d \psi}{d \theta} \right)^2 \right]$$
$$+ (1 + 2b_2 + b_1 \theta) \frac{d^{r+1} \psi}{d \theta^{r+1}} + \binom{r}{1} b_1 \frac{d^r \psi}{d \theta^r} = 0$$

$$\Rightarrow \quad b_2 \theta \left[\frac{d^{r+2} \psi}{d \theta^{r+2}} + \left\{ \frac{d^{r-1}}{d \theta^{r-1}} \left[2 \frac{d \psi}{d \theta} \cdot \frac{d^2 \psi}{d \theta^2} \right] \right\} \right]$$
$$+ rb_2 \left[\frac{d^{r+1} \psi}{d \theta^{r+1}} + \frac{d^{r-2}}{d \theta^{r-2}} \left\{ 2 \frac{d \psi}{d \theta} \cdot \frac{d^2 \psi}{d \theta^2} \right\} \right]$$
$$+ (1 + 2b_2 + b_1 \theta) \frac{d^{r+1} \psi}{d \theta^{r+1}} + rb_1 \frac{d^r \psi}{d \theta^r} = 0$$

Putting $\theta = 0$ and using the relation

$$\left[\frac{d^n \psi}{d\theta^n}\right]_{\theta=0} = \kappa_n, \text{ we get}$$

$$\left\{1 + (r+2)\,b_2\right\}\kappa_{r+1} + \binom{r}{1} b_1 \kappa_r + r b_2 \left\{\binom{r-1}{1}\kappa_2 . \kappa_{r-1}\right.$$

$$\left. + \binom{r-1}{2}\kappa_3 \kappa_{r-2} + \dots + \binom{r-1}{r-2}\kappa_{r-1}\kappa_2\right\} = 0 \text{ (On simplification)}$$

EXERCISE 8(i)

2. Derive the differential equation

$$\frac{1}{y}\cdot\frac{dy}{dx} = \frac{x+a}{b_0 + b_1 x + b_2 x^2}$$

as the limiting form of the hypergeometric distribution.

Show that, for the Pearsonian family of distributions :

$$\frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{(5\,\beta_2 - 6\,\beta_1 - 9)}$$

2. (*a*) State the differential equation for the Pearsonian system of curves and obtain the expressions for the constants in terms of moments. Obtain Type 1 distribution as a particular case of Pearsons's sytem of frequency curves and describe method of fitting it by moments.

(*b*) Describe the procedure for classifying the Pearson family of distributions into various types. Show that all Pearsonian distributions are determined by the first four moments.

Show that 'Normal', 'Beta' and 'Gamma' distributions belong to the Pearson family.

(*c*) Assign the following distribution to one of the Pearson's types. Give the reasons for your answers

(*i*) $dF = K\,e^{-x^2/2}\,(x^2)^{(n/2)-1}\,dx^2,\ 0 < x^2 < \infty$

(*ii*) $dF = K\left(1 + \dfrac{t^2}{n}\right)^{\left(\frac{-n-1}{2}\right)} dt,\ -\infty < t < \infty.$

3. What are the reasons for the adoption of the following general form to describe the Pearsonian system of frequency curves

$$\frac{d}{dx}f(x) = \frac{(x-a)\,f(x)}{b_0 + b_1 x + b_2 x^2}\ ?$$

Show that the Pearsonian curves can be characterised by a single criterion K. Outline the various types of curves for different values of K.

4. Obtain Pearson Type III curve in its usual form with mode as origin, from the basic differential equation of the Pearsonian system of curves and establish a method of fitting this curve to the given data by the method of moments.

Hence of otherwise, show that for this distribution $2\beta_2 = 3(\beta_1 + 2)$.

5. Derive the Beta distribution as a special case of the Pearsonian system of frequency functions expressed by

$$\frac{d(\log f)}{dx} = \frac{x+a}{b_0 + b_1 x + b_2 x^2}$$

6. (*a*) Derive Type 1 Pearsonian frequency curve and examine if the distribution given by

$$dP = y_0 (1 - y^2)^{(n-4)/2}\, dy, \; -1 \le y \le 1$$

reduces to that distribution.

(*b*) Express the constants y_0, a and m of the distribution :

$$f(x) = y_0 \left(1 - \frac{x^2}{a^2}\right)^m, \; -a < x < a$$

in terms of its μ_2 and β_2.

(*c*) Show that normal, gamma and beta distributions belong to the Pearsonian system.

7. Show that the following are memebers of the Pearson's sytem of curves and sketch them for some typical values of the constants.

(*i*) $f(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma} \cdot \exp\left(-\dfrac{1}{2} \cdot \dfrac{x^2}{\sigma^2}\right), \; -\infty < x < \infty$

(*iii*) $f(x) = \dfrac{1}{a\,B(\frac{1}{2}, m+1)} \left(1 - \dfrac{x^2}{a^2}\right)^m, \; -a \le x \le a$

(*iv*) $f(x) = \dfrac{1}{a\,B(\frac{1}{2}, m-1)} \left(1 + \dfrac{x^2}{a^2}\right)^m, \; -a \le x \le a,$

8. Show that the Pearsonian Type VI curve may be written

$$y = y_0 \left(1 - \frac{x^2}{a^2}\right)^{-m} \exp\left\{-v \tanh^{-1} \frac{x}{a}\right\}$$

and discuss its relationship with Type IV curve.

9. Show that for Pearson distribution :

$$\frac{d}{dx}(\log f) = \frac{x}{B_0 + B_1 x + B_2 x^2}$$

the range is unlimited in both the directions if $B_0 + B_1 x + B_2 x^2$ has no real roots, limited in one direction if roots are real and of the same sign, and limited in both directions if the roots are real and of opposite sign.

10. Investigate the properties and shapes which may be assumed by the frequency curve $y = f(x)$ which has the differential equation

$$\frac{1}{y} \cdot \frac{dy}{dx} = -\frac{2mx}{a^2 - x^2}$$

and obtain the probability integral.

Hint. $\frac{d}{dx}(\log y) = -\frac{2mx}{a^2 - x^2} \Rightarrow \log y = m \log(a^2 - x^2) + \log C$

$$\therefore \qquad y = k\left(1 - \frac{x^2}{a^2}\right)^m ; -a \le x \le a$$

which is type II distribution.

11. A family of distributions is defined by

$$\frac{1}{f} \cdot \frac{df}{dx} = \frac{x}{b_0 + b_2 x^2 + b_4 x^4}$$

and the frequency function $f = f(x)$ vanishes at the terminals of its range. Show that the moments about the mean are given by

$$b_0 (2s + 1) \mu_{2s} + b_2 (2s + 3) \mu_{2s+2} + b_4 (2s + 5) \mu_{2s+4} = -\mu_{2s+2}$$

8·13. Variate Transformations. Let T by any statistic which is asymptotically normally distributed with mean θ and variance $\psi(\theta)$, where $\psi(\theta)$ is some function of the parameter θ *i.e.*, $T \sim N(\theta, \psi(\theta))$. Let us transform T by a function g as $g(T)$. where g is a function which possesses first order derivative which is continuous and $g'(\theta) \neq 0$, where ($'$) denotes differentiation w.r.to. the parameter θ. Then $g(T)$ is normally distributed about mean $g(\theta)$ and variance $[g'(\theta)]^2 \cdot \psi(\theta)$, *i.e.*,

$$g(T) \sim N[g(\theta), \{g'(\theta)\}^2 \psi(\theta)], \qquad \text{...(8·47)}$$

asymptotically, provided $g'(\theta) \neq 0$ is continuous in the neighbourhood of θ.

In general Var $(T) = \psi(\theta)$ will be dependent on the parameter θ. We are interested in obtaining a function g such that the asymptotic variance of the transformed statistic $g(T)$ is independent of θ, *i.e.*, it should be constant. In other words, we want g such that

$$\text{Var}[g(T)] = [g'(\theta)]^2 \cdot \psi(\theta) = \text{constant} = c^2, \text{ (say)}$$

i.e., $$[g'(\theta)] = \frac{c}{\sqrt{\psi(\theta)}}$$

Integrating both sides w.r.to. θ, we get

$$g(\theta) = \int \frac{c}{\sqrt{\psi(\theta)}} d\theta \qquad \text{...(8·48)}$$

8·13·1. Uses of Variate Transformations. As discussed above, when we transform a statistic T to a function $g(T)$, the distribution of $g(T)$ is approximately normal and its asymptotic variance is independent of the population parameter θ. Hence the use of statistic $g(T)$ gives better results and confidence intervals than the original statistic T. The commonly used transformations are :

(1) Square Root Transformation.
(2) Sine Inverse or $\sin^{-1}$ Transformation.
(3) Logarithmic Transformation.
(4) Fisher's Z–Transformation.

In the following sections we shall discuss these transformations briefly.

8·13·2. Square Root Transformation. Square root transformation is a transformation for the Poisson variate. If a variable X follows Poisson distribution with parameter λ (assumed to be large), then we know that asymptotic distribution of X is normal as $(\lambda \to \infty)$ with $E(X) = \lambda$ and $\text{Var}(X) = \lambda = \psi(\lambda)$, in the above notations. Then (8·48) gives the function

$$g(\lambda) = \int \frac{c}{\sqrt{\lambda}} d\lambda = 2c\sqrt{\lambda}\,.$$

Now we select c in such a way that $2c = 1$

i.e., $c = 1/2$ so that $g(\lambda) = \sqrt{\lambda}$.

Hence the transformed variable is $g(X) = \sqrt{X}$. Using (8·47), the transformed variable $\sqrt{X}$ has the mean

$$g(\lambda) = \sqrt{\lambda}$$

$$\begin{aligned} \text{and } \text{Var}(\sqrt{X}) &= [g'(\lambda)]^2 \cdot \psi(\lambda) \\ &= \left(\frac{1}{2\sqrt{\lambda}}\right)^2 \cdot \lambda \\ &= 1/4\,. \end{aligned}$$

or alternatively

$$\text{Var}(\sqrt{X}) = \text{constant} = c^2 = (1/2)^2 = 1/4\,.$$

Hence $\sqrt{X} \sim N(\sqrt{\lambda}, 1/4)$, asymptotically. ...(8·49)

Ancombi has suggested the transformation $\sqrt{X+b}$, where b is a constant suitably chosen.

8·13·3. Sine Inverse or $\sin^{-1}$ Transformation. Sine-inverse is the transformation for stablizing the variance of a binomial variate. If p is the observed proportion of successes in a series of n independent trials with constant probability P of success for each trial then we know that the asymptotic distribution of p is asymptotically normal (as $n \to \infty$) with $E(p) = P$ and

$$\text{Var}(p) = PQ/n, \quad Q = 1 - P$$

i.e., $$p \sim N\left(P, \frac{PQ}{n}\right), \text{ as } n \to \infty.$$

In the usual notations

$$\psi(P) = \frac{PQ}{n} = \frac{P(1-P)}{n}\,.$$

Using (8·48), the transforming function

$$g(P) = \int \frac{c}{\sqrt{\psi(P)}}\, dP = c\sqrt{n} \int \frac{dP}{\sqrt{P(1-P)}}$$
$$= 2c\sqrt{n} \sin^{-1}(\sqrt{P})$$
$$\left[\because \frac{d}{dP} \sin^{-1}(\sqrt{P}) = \frac{1}{2\sqrt{P}.\sqrt{1-P}} \right]$$

Choosing the constant c so that

$$2c\sqrt{n} = 1 \quad \Rightarrow \quad c = \frac{1}{2\sqrt{n}},$$

we get

$$g(P) = \sin^{-1}(\sqrt{P})$$

Hence the transformed statistic is $g(p) = \sin^{-1}\sqrt{p}$. Using (8·47), $g(p)$ has mean $\sin^{-1}\sqrt{P}$ and

$$\text{Var}(\sin^{-1}\sqrt{p}) = [g'(P)]^2 . \psi(P)$$
$$= \left[\frac{1}{2\sqrt{P}\sqrt{1-P}}\right]^2 \times \frac{P(1-P)}{n}$$
$$= \frac{1}{4n}$$

or $\text{Var}(\sin^{-1}\sqrt{p}) = c^2 = \left(\frac{1}{2\sqrt{n}}\right)^2 = \frac{1}{4n} = \text{constant.}$

Hence $\quad \sin^{-1}\sqrt{p} \sim N\left(\sin^{-1}\sqrt{P}, \frac{1}{4n}\right)$, asymptotiically. ...(8·50)

If r is the observed number of successes in n trials so that $p = r/n$, then Ancombi has suggested that instead of $\sin^{-1}\sqrt{p} = \sin^{-1}\sqrt{r/n}$, the transformation should be $\sin^{-1}\sqrt{\frac{r+3/8}{n+3/8}}$.

8·13·4. Logarithmic Transformation. Log transformation is the transformation for stabilizing the variance of the distribution of sample variance. If s^2 is the sample variance in a simple of size n from normal population with variance σ^2, then the sampling distribution of s^2 is asymptotically normal (as $n \to \infty$) with

$$E(s^2) = \sigma^2 \quad \text{and} \quad \text{var}(s^2) = \frac{2\sigma^4}{n} \text{ (for large } n\text{)},$$

[*c.f.* Remark to Theorem 13·5]. In the usual notations we have $\psi(\sigma^2) = 2\sigma^4/n$. Using (8·48), the transforming function is

$$g(\sigma^2) = \int \frac{c\sqrt{n}}{\sqrt{2}\,\sigma^2}\, d\sigma^2 = \frac{c\sqrt{n}}{\sqrt{2}} \log \sigma^2.$$

We select c in such a way that

$$\frac{c\sqrt{n}}{\sqrt{2}} = 1 \quad \Rightarrow \quad c = \frac{\sqrt{2}}{\sqrt{n}},$$

so that $\quad g(\sigma^2) = \log_e \sigma^2.$

Hence the transformation for the statistic s^2 is $g(s^2) = \log s^2$ and using (8·47) the transformed statistic is normally distributed with mean

$$g(\sigma^2) \doteq \log \sigma^2 \text{ and}$$

$$\text{Var}[g(s^2)] = \{g'(\sigma^2)\}^2 \cdot \psi(\sigma^2)$$

$$= \left(\frac{1}{\sigma^2}\right)^2 \cdot \frac{2\sigma^4}{n}$$

$$= \frac{2}{n}$$

or $$\text{Var}[g(s^2)] = c^2 = \left(\frac{\sqrt{2}}{\sqrt{n}}\right)^2 = \frac{2}{n}.$$

Hence $$\log_e s^2 \sim N\left(\log_e \sigma^2, \frac{2}{n}\right), \text{ for large } n. \qquad ...(8\cdot51)$$

8·13·5. Fisher's z-Transformation. This transformation is suggested for stablizing the varaince of sampling distribution of correlation coefficient (*c.f.* chapter 11). If r is the sample correlation coefficient in sampling from a correlated bivariate normal population with correlation coefficient ρ then the asymptotic distribution of r, as $n \to \infty$ is normal with $E(r) = \rho$ and

$$\text{Var}(r) = \frac{(1-\rho^2)^2}{n} = \psi(\rho) \text{ for large } n. \text{ Using (8·47), we get}$$

$$g(\rho) = \int \frac{\sqrt{n}\,c}{1-\rho^2}\,d\rho = \frac{\sqrt{n}\,c}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$$

We select c in such a way that

$$\sqrt{n}\,c = 1 \quad \Rightarrow \quad c = 1/\sqrt{n},$$

so that $$g(\rho) = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right).$$

Hence using (8·47), the transformed statistic

$g(r) = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$, which is denoted by Z, is normally distributed with mean

$$g(\rho) = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right) \text{ and}$$

$$\text{Var}[g(r)] = c^2 = \frac{1}{n}$$

or $$\text{Var}[g(r)] = [g'(\rho)]^2\,\psi(\rho)$$

$$= \left[\frac{1}{1-\rho^2}\right]^2 \cdot \frac{(1-\rho^2)^2}{n}$$

$$= \frac{1}{n}, \text{ for large } n$$

Hence $$Z = \frac{1}{2} \log_e \left[\frac{1+r}{1-r} \right] \sim N \left[\frac{1}{2} \log_e \frac{1+\rho}{1-\rho}, \frac{1}{n} \right], \text{ for large } n. \quad \ldots(8\cdot52)$$

Prof. R.A. Fisher proved that the transformed statistic $Z = g(r)$ is normally distributed even if n is small and that for exact samples (small n),

$$Z \sim N \left[\frac{1}{2} \log_e \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right].$$

For the various applications of this transformation the reader is referred to § 14·7·2.

Remark. We have : $Z = \frac{1}{2} \log_e \left[\frac{1+r}{1-r} \right] = \tanh^{-1}(r)$

Hence Z–transformation is also called the tan-hyperbolic-inverse transformation.

8·14 Order Statistics Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed variates, each with cumulative distribution function $F(x)$. If these variables are arranged in ascending order of magnitude and then written as $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, we call $X_{(r)}$ as the rth order statistic, $r = 1, 2, \ldots, n$. The $X_{(r)}$'s because of the inequality relations among them are necessarily dependent.

Remark. If we write these ordered values as

$Y_1 \leq Y_2 \leq \ldots \leq Y_n$, then:

$Y_r = X_{(r)} =$ rth smallest of $X_1, X_2, \ldots, X_n$

$Y_1 = X_{(1)} =$ The smallest of $X_1, X_2, \ldots, X_n$

$Y_n = X_{(n)} =$ The largest of $X_1, X_2, \ldots, X_n$

8·14·1. Cumulative Distribution Function of a Single Order Statistic. Let $F_r(x)$, $r = 1, 2, \ldots, n$ denote the c.d.f. of the rth order statistic $X_{(r)}$. Then the c.d.f. of the largest order statistic $X_{(n)}$ is given by :

$$\begin{aligned} F_n(x) &= P(X_{(n)} \leq x) = P(X_i \leq x \,;\, i = 1, 2, \ldots, n) \\ &= P(X_1 \leq x \cap X_2 \leq x \cap \ldots \cap X_n \leq x) \\ &= P(X_1 \leq x) \,.\, P(X_2 \leq x) \ldots P(X_n \leq x) \qquad (\because X_i\text{'s are independent}) \\ &= [F(x)]^n, \qquad \ldots(8\cdot53) \end{aligned}$$

since $X_1, X_2, \ldots, X_n$ are identically distributed.

The c.d.f. of the smallest order statistic $X_{(1)}$ is given by :

$$\begin{aligned} F_1(x) &= P(X_{(1)} \leq x) \\ &= 1 - P(X_{(1)} > x) \\ &= 1 - P[X_i > x \,;\, i = 1, 2, \ldots, n] \\ &= 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - \prod_{i=1}^{n} [1 - P(X_i \leq x)] \\ &= 1 - [1 - F(x)]^n, \qquad \ldots(8\cdot54) \end{aligned}$$

since $X_1, X_2, \ldots, X_n$ are i.i.d. rv's.

In general, the c.d.f. of the rth order statistic $X_{(r)}$ is given by:

$$F_r(x) = P(X_{(r)} \leq x)$$
$$= P \text{ [At least } r \text{ of the } X_i \text{ 's are } \leq x]$$
$$= \sum_{j=r}^{n} P \text{ [Exactly } j \text{ of the } n, X_i \text{ 's are } \leq x],$$
$$= \sum_{j=r}^{n} \binom{n}{j} F^j(x)\,[1-F(x)]^{n-j}, \qquad ...(8·55)$$

by using Binomial probability model.

Remarks. 1. (8·55) can also be written as [See Remark 2 to Example 7·23]:

$$F_r(x) = I_{F(x)}\,(r, n-r+1), \qquad ...(8·56)$$

where
$$I_p\,(a, b) = \frac{1}{\beta\,(a, b)} \int_0^p t^{a-1}(1-t)^{b-1}\,dt \qquad ...(8·56\,a)$$

is the '*incomplete Beta Function*' and has been tabulated in Biometrika tables by Pearson and Hartley.

(8·56) and (8·56 a) show that the probability points of an order statistic can be obtained with the help of incomplete beta function.

2. Taking $r = 1$ and $r = n$ in (8·55), we get respectively:

$$F_1(x) = \sum_{j=1}^{n} \binom{n}{j} F^j\,(x)\,[1-F\,(x)]^{n-j}$$
$$= 1 - \left\{\binom{n}{j} F^j\,(x)\,[1-F\,(x)]^{n-j}\right\}_{j=0}$$
$$= 1 - [1 - F\,(x)]^n \qquad ...(8·56\,b]$$

and
$$F_n\,(x) = F^n\,(x), \qquad ...(8·56\,c)$$

the results which have already been obtained in (8·54) and (8·53) respectively,

8·14·2. Probability Density Function (p.d.f.) of a Single Order Statistic. The results in (8·53) to (8·55) are valid for both discrete and continuous r.v.'s. We shall now assume that X_i's are i.i.d. continuous r.v.'s with p.d.f. $f(x) = F'(x)$. If $f_r\,(x)$ denotes the p.d.f. of $X_{(r)}$ then from (8·55) or (8·56) we get:

$$f_r\,(x) = \frac{d}{dx}[F_r\,(x)] = \frac{d}{dx}[I_{F(x)}\,(r, n-r+1)]$$
$$= \frac{d}{dx}\left[\frac{1}{\beta(r, n-r+1)} \int_0^{F(x)} t^{r-1}\,(1-t)^{n-r}\,dt\right] \qquad ...(8·57)$$

Let us write

$$g(t) = \int t^{r-1}\,(1-t)^{n-r}\,dt \Rightarrow g'\,(t) = t^{r-1}\,(1-t)^{n-r} \qquad ...(*)$$

$$\Rightarrow \int_0^{F(x)} t^{r-1}(1-t)^{n-r}\,dt = \left| g(t) \right|_0^{F(x)} = g(F(x)) - g(0)$$

$$\Rightarrow \frac{d}{dx}\int_0^{F(x)} t^{r-1}(1-t)^{n-r}\,dt = g'(F(x))\,.\,f(x) \qquad (\because g(0) \text{ is constant})$$

$$= [F(x)]^{r-1}[1-F(x)]^{n-r} f(x) \quad [\text{ Using } (*)]$$

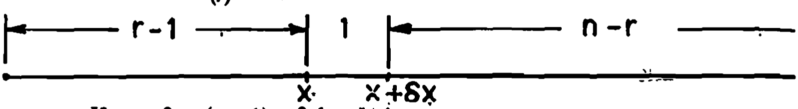
Substituting in (8·57) we get :

$$f_r(x) = \frac{1}{\beta(r, n-r+1)}\,.\,F^{r-1}(x)[1-F(x)]^{n-r}\,.\,f(x) \qquad \ldots(8·58)$$

Aliter. By definition of a p.d.f. we get:

$$f_r(x) = \lim_{\delta x \to 0} \frac{P[x < X_{(r)} \le x + \delta x]}{\delta x} \qquad \ldots(8·59)$$

The event E: $x < X_{(r)} \le x + \delta x$ can materialise as follows:



$X_i \le x$ for $(r-1)$ of the X_i's

and $x < X_i \le x + \delta x$ for one X_i

and $X_i \ge x + \delta x$ for the remaining $(n-r)$ of the X_i's.

Hence by the multinomial probability law we have:

$$P(x < X_{(r)} \le x + \delta x) = \frac{n!}{(r-1)!\,1!\,(n-r)!} p_1^{r-1}\,.\,p_2^1\,.\,p_3^{n-r} \qquad \ldots(8·60)$$

where $p_1 = P(X_i \le x) = F(x)$

$p_2 = P(x < X_i \le x + \delta x) = F(x + \delta x) - F(x)$

and $p_3 = P(X_i \ge x + \delta x) = 1 - P(X_i \le x + \delta x) = 1 - F(x + \delta x)$

Substituting in (8·60), we get:

$$f_r(x) = \lim_{\delta x \to 0} \frac{P(x < X_{(r)} \le x + \delta x)}{\delta x}$$

$$= \frac{1}{\beta(r, n-r+1)} \times F^{r-1}(x) \times \lim_{\delta x \to 0}\left[\frac{F(x+\delta x) - F(x)}{\delta x}\right] \times \lim_{\delta x \to 0}[1 - F(x+\delta x)]^{n-r}$$

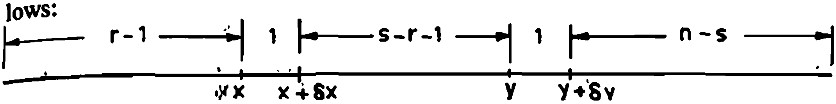
$$= \frac{1}{\beta(r, n-r+1)}\,.\,F^{r-1}(x)\,.\,f(x)\,.\,[1-F(x)]^{n-r},$$

as in (8·58).

8·14·3. Joint p.d.f. of two Order Statistics. Let us denote the joint p.d.f. of $X_{(r)}$ and $X_{(s)}$, where $1 \le r < s \le n$ by $f_{rs}(x, y)$. Then,

$$f_{rs}(x, y) = \lim_{\substack{\delta x \to 0 \\ \delta y \to 0}} \frac{P\left[x \le X_{(r)} \le x + \delta x \cap y \le X_{(s)} \le y + \delta y\right]}{\delta x\, \delta y} \qquad \ldots(8·6)$$

The event $E = \left\{ x \le X_{(r)} \le x + \delta x \cap y \le X_{(s)} \le y = \delta y \right\}$ can materialise as follows:



$X_i \le x$ for $r-1$ of the X_i's,
$x < X_i \le x + \delta x$ for one X_i,
$x + \delta x < X_i \le y$ for $(s-r-1)$ of X_i's,
$y < X_i \le y + \delta y$ for one X_i.
and $X_i > y + \delta y$ for $(n-s)$ of the X_i's

Hence using multinomial probability law, we get

$$P(E) = P\left[x \le X_{(r)} \le x + \delta x \cap y \le X_{(s)} < y + \delta y \right]$$
$$= \frac{n!}{(r-1)!\,1!\,(s-r-1)!\,1!\,(n-s)!} p_1^{r-1} \cdot p_2 p_3^{s-r-1} \cdot p_4 p_5^{n-s} \quad \ldots(8\cdot62)$$

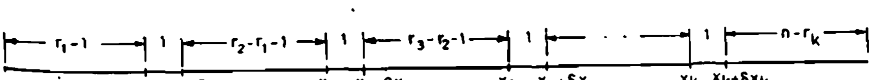
where

$$p_1 = P(X_i \le x) = F(x)$$
$$p_2 = P(x < X_i \le x + \delta x) = F(x + \delta x) - F(x)$$
$$p_3 = P(x + \delta x < X_i \le y) = F(y) - F(x + \delta x)$$
$$p_4 = P(y < X_i \le y + \delta y) = F(y + \delta y) - F(y)$$
$$p_5 = P(X_i > y + \delta y) = 1 - P(X_i \le y + \delta y) = 1 - F(y + \delta y)$$

Substituting in (8·62) and using (8·61) we get:

$$f_{rs}(x,y) = \lim_{\substack{\delta x \to 0 \\ \delta y \to 0}} \frac{P(E)}{\delta x\, \delta y}$$
$$= \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} \times F^{r-1}(x) \times \lim_{\delta x \to 0} \frac{[F(x+\delta x) - F(x)]}{\delta x}$$
$$\times \lim_{\delta y \to 0} \left[\frac{F(y+\delta y) - F(y)}{\delta y} \right] \times \lim_{\delta y \to 0} [1 - F(y + \delta y)]^{n-s}$$
$$\times \lim_{\delta x \to 0} [F(y) - F(x + \delta x)]^{s-r-1}$$
$$= \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} F^{r-1}(x) \cdot f(x) \cdot [F(y) - F(x)]^{s-r-1} f(y) \cdot [1 - F(y)]^{n-s} \quad \ldots(8\cdot63)$$

8·14·4 Joint p.d.f. of k – Order Statistics. The joint p.d.f. of k – order statistics $X_{(r_1)}, X_{(r_2)}, \ldots, X_{(r_k)}$ where $1 \le r_1 < r_2 < \ldots < r_k \le n$ and $1 \le k \le n$ is for $x_1 \le x_2 \le \ldots \le x_k$ given by [on using the following configuration and the multinomial probability law as in § 8·14·3]:

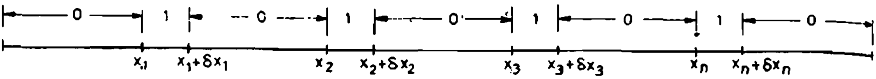


$$f_{r_1, r_2, \dots, r_k}(x_1, x_2, \dots, x_k) = \frac{n!}{(r_1 - 1)!\,(r_2 - r_1 - 1)! \dots (r_k - r_{k-1} - 1)!\,(n - r_k)!}$$
$$\times F^{r_1 - 1}(x_1) \times f(x_1) \times \left[F(x_2) - F(x_1) \right]^{r_2 - r_1 - 1} \times f(x_2)$$
$$\times \left[F(x_3) - F(x_2) \right]^{r_3 - r_2 - 1} \times f(x_3) \times \dots \times f(x_k) \left[1 - F(x_k) \right]^{n - r_k} \quad \dots(8\cdot64)$$

8·14·5. Joint p.d.f. of all n - Order Statistics. In particular the joint p.d.f. of all the n order statistics is obtained on taking $k = n$ in (8·64). This implies that $r_i = i$ for $i = 1, 2, \dots, n$. Hence joint.p.d.f. of $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ is given by:

$$f_{1, 2, \dots, n}(x_1, x_2, \dots, x_n) = n!\, f(x_1) f(x_2) \dots f(x_n) \quad \dots(8\cdot65)$$

Aliter. We can easily obtain (8·65) by using the following configuration:



and the multinomial probability law as in § 8·14·3.

8·14·6. Distribution of Range and other Systematic Statistics. Let us obtain the p.d.f. the statistic $W_{rs} = X_{(s)} - X_{(r)}$; $r < s$. We start with the joint p.d.f. of $X_{(r)}$ and $X_{(s)}$ given in (8·63) and transform $[X_{(r)}, X_{(s)}]$ to the new variables W_{rs} and $X_{(r)}$ s.t.

$$w_{rs} = y - x\ ; \quad x = x \quad \text{s.t.;} \quad y = x + w_{rs} \text{ and } x = x$$

$$\therefore \quad J = \frac{\partial(x, y)}{\partial(x, w_{rs})} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1 \quad \Rightarrow \quad |J| = 1$$

The joint p.d.f. $f_{rs}(x, y)$ in (8·63) transforms to the joint p.d.f. of $X_{(r)}$ and W_{rs} as given below:

$$g(x, w_{rs}) = c_{rs} \cdot F^{r-1}(x) \cdot f(x) \left[F(x + w_{rs}) - F(x) \right]^{s-r-1}$$
$$\times f(x + w_{rs}) \times \left[1 - F(x + w_{rs}) \right]^{n-s} \quad \dots(8\cdot66)$$

$$\text{where } c_{rs} = \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} \quad \dots(8\cdot67)$$

Integrating (8·66) w.r.to. x from $-\infty$ to ∞, we obtain the p.d.f. of W_{rs} as:

$$g(w_{rs}) = c_{rs} \int_{-\infty}^{\infty} \left\{ F^{r-1}(x) f(x) \left[F(x + w_{rs}) - F(x) \right]^{s-r-1} \cdot f(x + w_{rs}) \right.$$
$$\left. \cdot \left[1 - F(x + w_{rs}) \right]^{n-s} \right\} dx \quad \dots(8\cdot68)$$

Remark. Distribution of Range $W = X_{(n)} - X_{(1)}$. Taking $r = 1$ and $s = n$ in (8·68), we obtain the p.d.f. of the range $W = X_{(n)} - X_{(1)}$ as:

$$g(w) = n(n-1) \int_{-\infty}^{\infty} f(x) \left[F(x + w) - F(x) \right]^{n-2} \cdot f(x + w)\, dx\ ; \ w \geq 0 \quad \dots(8\cdot69)$$

The c.d.f of W is rather simple as given below:

$$G(w) = P(W \le w) = \int_0^w g(u)\, du$$

$$= \int_0^w \left\{ n(n-1) \int_{-\infty}^{\infty} f(x) [F(x+u) - F(x)]^{n-2} f(x+u)\, dx \right\} du$$

$$= n \int_{-\infty}^{\infty} \left[f(x) \left\{ \int_0^w (n-1) f(x+u) (F(x+u) - F(x))^{n-2}\, du \right\} \right] dx$$

$$= n \int_{-\infty}^{\infty} f(x) [F(x+w) - F(x)]^{n-1}\, dx$$

Example 8·48 *Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with continuous density. Show that $Y_1 = \min(X_1, X_2, \ldots, X_n)$, is exponential with parameter $n\lambda$ if and only if each X_i is exponential with parameter λ.*

Solution. Let X_i be i.i.d. exponential variates with parameter λ and p.d.f.

$$f(x) = \lambda e^{-\lambda x};\ x \ge 0,\ \lambda > 0 \qquad \ldots(i)$$

$$F(x) = P(X \le x) = \int_0^x f(u)\, du = \lambda \int_0^x e^{-\lambda u}\, du = 1 - e^{-\lambda x} \qquad \ldots(ii)$$

Distribution function $G(\,.\,)$ of $Y_1 = \min(X_1, X_2, \ldots, X_n)$ is given by:

$$G_{Y_1}(y) = P(Y_1 \le y) = 1 - [1 - F_1(y)]^n \qquad \text{[From (8·54)]}$$

$$= 1 - \left[1 - (1 - e^{-\lambda y}) \right]^n = 1 - e^{-n\lambda y} \qquad \ldots(iii)$$

[From (*ii*)]

which is the distribution function of exponential distribution with parameter $n\lambda$. Hence $Y_1 = \min(X_1, X_2, \ldots, X_n)$, has exponential distribution with parameter $n\lambda$.

Conversely, Let $Y_1 = \min(X_1, X_2, \ldots, X_n) \sim \text{Exp}(n\lambda)$ so that

$$P(Y_1 \le y) = 1 - e^{-n\lambda y} \quad \Rightarrow \quad P(Y_1 \ge y) = e^{-n\lambda y} \qquad \ldots(iv)$$

$$\Rightarrow \quad P\left[\min(X_1, X_2, \ldots, X_n) \ge y \right] = e^{-n\lambda y}$$

$$\Rightarrow P\left[(X_1 \ge y) \cap (X_2 \ge y) \cap \ldots (X_n \ge y) \right] = e^{-n\lambda y}$$

$$\Rightarrow \quad \prod_{i=1}^{n} P(X_i \ge y) = e^{-n\lambda y}$$

$$\left[P(X_i \ge y) \right]^n = e^{-n\lambda y} \qquad [\because X_i\text{'s are } i.i.d.]$$

$$\Rightarrow \quad P(X_i \ge y) = e^{-\lambda y}$$

$$\Rightarrow \qquad P(X_i \le y) = 1 - e^{-\lambda y}$$

which is the distribution function of Exp (λ) distribution. Hence X_i's are i.i.d. Exp (λ).

Example 8·49. *For the exponential distribution* $f(x) = e^{-x}, x \ge 0$; *show that the cumulative distribution function (c.d.f.) of* $X_{(n)}$ *in a random sample of size* n *is* $F_n(x) = (1 - e^{-x})^n$. *Hence prove that as* $n \to \infty$, *the c.d.f. of* $X_{(n)} - \log n$ *tends to the limiting form* $\exp[-(\exp(-x))], -\infty < x < \infty$.

Solution. Here $f(x) = e^{-x}, x \ge 0$; $F(x) = P(X \le x) = 1 - e^{-x}$ ...(*)

The c.d.f. of $X_{(n)}$ is given by [From (8·53)]

$$F_n(x) = P[X_{(n)} \le x] = [F(x)]^n = (1 - e^{-x})^n \qquad [\text{From } (*)] \quad \ldots(**)$$

The c.d.f. $G_n(\,.\,)$ of $X_{(n)} - \log n$ is given by:

$$G_n(x) = P\left[X_{(n)} - \log n \le x\right]$$

$$= P[X_{(n)} \le x + \log n]$$

$$= \left[1 - e^{-(x + \log n)}\right]^n \qquad [\text{From } (**)]$$

$$= \left[1 - \frac{e^{-x}}{n}\right]^n \qquad \left[\because e^{-\log n} = e^{\log n^{-1}} = \frac{1}{n}\right]$$

$$\therefore \quad \lim_{n \to \infty} G_n(x) = \lim_{n \to \infty} \left[1 - \frac{e^{-x}}{n}\right]^n = \exp\left[-e^{-x}\right]$$

$$\left[\because \lim_{n \to \infty} \left(1 + \frac{m}{n}\right)^n = e^m\right]$$

Example 8·50 *Show that for a random sample of size 2 from* $N(0, \sigma^2)$ *population,* $E(X_{(1)}) = -\sigma/\sqrt{\pi}$ **[Delhi Univ. M.Sc. (Stat.), 1988, 1982]**

Solution. For $n = 2$, the p.d.f. $f_1(x)$ of $X_{(1)}$ is given by: [From (8·58)]

$$f_1(x) = \frac{1}{\beta(1, 2)}[1 - F(x)]f(x) = 2[1 - F(x)] \,.\, f(x)\,; -\infty < x < \infty$$

where $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/2\sigma^2}$ $\qquad \left[\because X \sim N(0, \sigma^2)\right]$

$$\therefore E(X_{(1)}) = \int_{-\infty}^{\infty} x \,.\, f_1(x)\, dx = 2\int_{-\infty}^{\infty} [1 - F(x)] \,.\, x f(x)\, dx \qquad \ldots(i)$$

We have: $\log f(x) = -\log(\sqrt{2\pi}\,\sigma) - \dfrac{x^2}{2\sigma^2}$

Differentiating w.r.t. x we get:

$$\frac{f'(x)}{f(x)} = -\frac{x}{\sigma^2}$$

$$\Rightarrow \qquad \int x f(x)\, dx = -\sigma^2 \int f'(x)\, dx = -\sigma^2 f(x) \qquad \ldots(ii)$$

Integrating (i) by parts and using (ii), we get:

$$E(X_{(1)}) = 2 \cdot \left\{ [1 - F(x)](-\sigma^2 f(x)) \right\}_{-\infty}^{\infty} - 2\int_{-\infty}^{\infty} (-\sigma^2 f(x))(-f(x))\, dx$$

$$= -2\sigma^2 \int_{-\infty}^{\infty} [f(x)]^2 dx = -\frac{1}{\pi} \int_{-\infty}^{\infty} e^{-x^2/\sigma^2} dx$$

$$= -\frac{1}{\pi} \cdot \frac{\sqrt{\pi}}{(1/\sigma)} \qquad \left(\because \int_{-\infty}^{\infty} e^{-a^2x^2} dx = \sqrt{\pi}/a \right)$$

$$= -\sigma/\sqrt{\pi}$$

Example 8·51. *Show that in odd samples of size n from $U[0, 1]$ population, the mean and variance of the distribution of median are $1/2$ and $1/[4(n+2)]$ respectively.*

Solution. We have: $\quad f(x) = 1 \; ; \; 0 \le x \le 1$

$$F(x) = P(X \le x) = \int_0^x f(u)\, du = \int_0^x 1 \,.\, du = x$$

Let $n = 2m + 1$ (odd), where m is a positive integer ≥ 1. Then median observation is $X_{(m+1)}$. Taking $r = (m + 1)$ in (8·58), the p.d.f of median $X_{(m+1)}$ is given by:

$$f_{m+1}(x) = \frac{1}{\beta(m+1, m+1)} \cdot x^m (1-x)^m$$

$$\therefore \quad E(X_{(m+1)}) = \frac{1}{\beta(m+1, m+1)} \cdot \int_0^1 x \,.\, x^m (1-x)^m dx$$

$$= \frac{\beta(m+2, m+1)}{\beta(m+1, m+1)}.$$

$$= \frac{\Gamma(m+2) \,.\, \Gamma(m+1)}{\Gamma(m+3)} \times \frac{\Gamma(2m+2)}{\Gamma(m+1)\Gamma(m+1)}$$

$$= \frac{m+1}{2m+2} = \frac{1}{2} \qquad \text{(On simplification)}$$

$$E(X^2_{(m+1)}) = \int_0^1 x^2 f_{m+1}(x)\, dx = \frac{1}{\beta(m+1, m+1)} \cdot \int_0^1 x^{m+2} (1-x)^m dx$$

$$= \frac{\beta(m+3, m+1)}{\beta(m+1, m+1)} = \frac{m+2}{2(2m+3)}$$

$$\therefore \quad \text{Var}(X_{(m+1)}) = E(X^2_{(m+1)}) - [E X_{(m+1)}]^2$$

$$= \frac{m+2}{2(2m+3)} - \frac{1}{4} = \frac{1}{4(2m+3)} = \frac{1}{4(n+2)}$$

Example 8·52. *Let $X_1, X_2, ..., X_n$ be i.i.d. non-negative random variables of the continuous type with p.d.f. $f(\,.\,)$ and distribution function $F(\,.\,)$.*

If $E|X| < \infty$, Show that $E|X_{(r)}| < \infty$.

(b) *Write* $M_n = X_{(n)} = \max(X_1, X_2, ..., X_n)$. *Show that*

$$E(M_n) = E(M_{n-1}) + \int_0^{\infty} F^{n-1}(x)\,[1 - F(x)]\,dx\,;\ n = 2, 3, ...$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

Hence evaluate $E(M_n)$ if $X_1, X_2, ..., X_n$ have common distribution function:

$$F(x) = x\,;\ 0 < x < 1.$$

Solution (a) $E\,|X_{(r)}| = \int_0^{\infty} |x| \,.\, f_r(x)\,dx$

$(\because X$ is non – negative continuous r.v.)

$$= \int_0^{\infty} |x| \,.\, \frac{n!}{(r-1)!\,(n-r)!} f(x)\,.\,F^{r-1}(x)\,[1 - F(x)]^{n-r}\,dx$$

$$\leq n\binom{n-1}{r-1} \,.\, \int_0^{\infty} |x|\,f(x)\,dx$$

$$\leq n\binom{n-1}{r-1} E\,|X|$$

Hence $E\,|X_{(r)}| < \infty$ if $E\,|X| < \infty$.

(b) The p.d.f. $f_n(x)$, of $M_n = X_{(n)}$ is given by:

$$f_n(x) = n\,[F(x)]^{n-1}\,.\,f(x)$$

$$E(M_n) = \lim_{a \to \infty} \int_0^{a} x\,f_n(x)\,dx \qquad (\because X \geq 0, a.s.)$$

$$\therefore\ E(M_n) = \lim_{a \to \infty} n \int_0^{a} x\,[F(x)]^{n-1}\,.\,f(x)\,dx$$

Integrating by parts we get:

$$E(M_n) = n\,.\lim_{a \to \infty} \left[\left| x\,.\,\frac{F^n(x)}{n} \right|_0^a - \int_0^a \frac{F^n(x)}{n}\,.\,1\,dx \right]$$

$$= \lim_{a \to \infty} \left[a\,F^n(a) - \int_0^a F^n(x)\,dx \right]$$

$$= \lim_{a \to \infty} \left[\int_0^a (1 - F^n(x) - 1)\,dx + a\,F^n(a) \right]$$

$$= \lim_{a \to \infty} \left[\int_0^a \left(1 - F^n(x)\right) dx - a + a F^n(a) \right]$$

$$= \lim_{a \to \infty} \left[\int_0^a \left(1 - F^n(x)\right) dx - a \left(1 - F^n(a)\right) \right]$$

Since $E M_n$ exists, (By part (a)),

$$a \cdot P(M_n > a) = a \left[1 - P(M_n \le a) \right] = a \left[1 - F^n(a) \right]$$

$$\longrightarrow 0 \text{ as } a \longrightarrow \infty.$$

$$\therefore E(M_n) = \lim_{a \to \infty} \int_0^a \left(1 - F^n(x)\right) dx = \int_0^\infty \left(1 - F^n(x)\right) dx \qquad \ldots(*)$$

$$= \int_0^\infty \left(1 - F^{n-1}(x) \cdot F(x)\right) dx$$

$$= \int_0^\infty \left[1 - F^{n-1}(x) \{ 1 - (1 - F(x)) \} \right] dx$$

$$= \int_0^\infty \left(1 - F^{n-1}(x)\right) dx + \int_0^\infty F^{n-1}(x) \left[1 - F(x) \right] dx$$

$$= E\left(M_{n-1}\right) + \int_0^\infty F^{n-1}(x) \left[1 - F(x) \right] dx \quad [\text{From } (*)] \qquad \ldots(**)$$

If $X \sim U[0, 1]$, then

$$f(x) = 1 \,;\; 0 < x < 1 \qquad \text{and} \qquad F(x) = x \,;\; 0 < x < 1$$

Substituting in (**), we get:

$$E(M_n) - E(M_{n-1}) = \int_0^1 x^{n-1} (1 - x)\, dx$$

$$\Rightarrow \qquad E(M_n) - E(M_{n-1}) = \frac{1}{n} - \frac{1}{n+1} \qquad \ldots(***)$$

Changing n to $n-1, n-2, \ldots, 2, 1$ in (***) we get respectively:

$$E(M_{n-1}) - E(M_{n-2}) = \frac{1}{n-1} - \frac{1}{n}$$

$$E(M_2) - E(M_1) = \frac{1}{2} - 1$$

$$E(M_1) - E(M_0) = 1 - \frac{1}{2}$$

Adding (***) and the above equations and noting that $E(M_0) = 0$, we get:

$$E(M_n) = 1 - \frac{1}{n+1} = \frac{n}{n+1}$$

Example 8·53 (*a*) *Find the p.d.f. of* $X_{(r)}$ *in a random sample of size* n *from the exponential distribution:*

$$f(x) = \alpha\, e^{-\alpha x},\ \alpha > 0,\ x \geq 0$$

(*b*) *Show that* $X_{(r)}$ *and* $W_{rs} = X_{(s)} - X_{(r)}$, $r < s$, *are independently distributed.*

(*c*) *What is the distribution of* $W_1 = X_{(r+1)} - X_{(r)}$?

Solution. Here $F(x) = P(X \leq x) = \int_0^x \alpha\,.\,e^{-\alpha u}\,du = 1 - e^{-\alpha x}$ ...(*i*)

The p.d.f. of $X_{(r)}$ is given by:

$$f_r(x) = \frac{1}{\beta(r, n-r+1)}\,.\,[F(x)]^{r-1}\,.\,[1 - F(x)]^{n-r}\,.\,f(x)$$

$$= \frac{1}{\beta(r, n-r+1)}\,.\,\left(1 - e^{-\alpha x}\right)^{r-1}\,.\,e^{-\alpha x(n-r)}\,.\,\alpha\,.\,e^{-\alpha x}$$

$$= \frac{1}{\beta(r, n-r+1)}\,.\,\alpha\,.\,e^{-\alpha x(n-r+1)}\,.\left[1 - e^{-\alpha x}\right]^{r-1};\ x > 0$$

(*b*) The joint p.d.f. of $X_{(r)}$ and $W_{rs} = X_{(s)} - X_{(r)}$ is given by [From (8·66)]

$$g(x, w_{rs}) = c_{rs}\,.\,F^{r-1}(x)\,f(x)\left[F(x + w_{rs}) - F(x)\right]^{s-r-1}$$

$$\times f(x + w_{rs})\left[1 - F(x + w_{rs})\right]^{n-s}$$

$$= \frac{n!}{(r-1)!\,(n-r)!} \times \frac{(n-r)!}{(s-r-1)!\,(n-s)!} \times \left[1 - e^{-\alpha x}\right]^{r-1}\ \alpha\, e^{-\alpha x}$$

$$\times \left[e^{-\alpha x} - e^{-\alpha(x + w_{rs})}\right]^{s-r-1} \times \alpha\, e^{-\alpha(x + w_{rs})} \times \left[e^{-\alpha(x + w_{rs})}\right]^{n-s}$$

$$= \left[\frac{1}{\beta(r, n-r+1)}\,.\,\alpha\, e^{-\alpha x(n-r+1)}\left(1 - e^{-\alpha x}\right)^{r-1}\right]$$

$$\times \left[\frac{1}{\beta(s-r, n-s+1)}\,.\,\alpha\,.\,e^{-(n-s+1)\alpha w_{rs}}\left(1 - e^{\alpha . w_{rs}}\right)^{s-r-1}\right] \quad ...(ii)$$

$\Rightarrow$ $X_{(r)}$ and W_{rs} are independently distributed.

(*c*) Taking $s = r + 1$ in (*ii*), the p.d.f. of $W_1 = X_{(r+1)} - X_{(r)}$ becomes:

$$g(w_1) = \frac{1}{\beta(1, n-r)}\,.\,\alpha\,.\,e^{-\alpha(n-r)w_1}$$

$$= (n-r)\,\alpha\,.\,e^{-(n-r)\alpha w_1};\ w_1 \geq 0$$

which shows that W_1 has an exponential distribution with parameter $(n-r)\,\alpha$.

EXERCISE 8 (j)

(1) (*a*) Obtain the distribution function and hence the p.d.f. of the rth order statistic $X_{(r)}$ in a random sample of size n from a population with continuous distribution function $P(\,.\,)$. Deduce the p.d.f.'s of the smallest and the largest sample observations. **[Delhi Univ. M.Sc. (Stat.), 1987]**

(*b*) Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population having continuous distribution function $P(x)$. Define the rth order statistic $X_{(r)}$ and obtain its distribution function and hence its p.d.f.

[Delhi Univ. M.Sc. (Stat.), 1983]

2. Define rth order statistic $X_{(r)}$. Obtain the joint p.d.f. of $X_{(r)}$ and $X_{(s)}$, $r < s$, in a random sample of size n from a population with continuous distribution function $P(\,.\,)$. Hence deduce the p.d.f. of sample range $W = X_{(n)} - X_{(1)}$.

[Delhi Univ. M.Sc. (Stat.), 1988, 1982]

3. Obtain the distribution function and hence the p.d.f. of the smallest sample abservation $X_{(1)}$ in a random sample of size n from a population with a continuous distribution function $F(x)$. Show that for random sample of size 2 from normal population $N(0, \sigma^2)$, $E(X_{(1)}) = -\sigma/\sqrt{\pi}$

[Bombay Univ. M.Sc. (Stat.), 1992]

4. Let $X_1, X_2, ..., X_n$ be n independent variates, X_i having a geometric distribution with parameter p_i, *i.e.*,

$P(X_i = x_i) = q_i^{x_i - 1} \cdot p_i\,;\ q_i = 1 - p_i, x_i = 1, 2, 3, \ldots$

Show that $X_{(1)}$ is distributed geometrically with parameter $(1 - q_1 q_2 \ldots q_n)$

[Delhi Univ. M.Sc. (Stat.), 1983]

(*b*) Let $X_1, X_2, ..., X_n$ be n independent variates, X_i having a geometric distribution with parameter p_i *i.e.*

$$P(X_i = x_i) = q_i^{x_i - 1} \cdot p_i\,;\ q_i = 1 - p_i, x_i = 1, 2, 3, ...,$$

Show that $X_{(1)}$ is distributed geometrically with parameter $(1 - q_1 q_2 q_3 \ldots q_n)$.

5. For a random sample of size n from a continuous population whose p.d.f. $p(x)$ is symmetrical at $x = \mu$, show that

$$f_r(\mu + x) = f_{n-r+1}(\mu - x),$$

where $f_r(\,.\,)$ is the p.d.f. of $X_{(r)}$.

Hint. $f(\mu + x) = f(\mu - x)$

$$F(\mu + x) = P(X \le \mu + x) = P(X \ge \mu - x) \qquad \text{(By symmetry)}$$
$$= 1 - P(X \le \mu - x) = 1 - F(\mu - x).$$

6. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population having continuous distribution function $F(x)$.

Define the order statistic of rank k, $1 \le k \le n$. Find its distribution function. Show that for the rectangular distribution

$$f(x) = 1/\theta_2, \quad \theta_1 - \tfrac{1}{2}\theta_2 \le x \le \theta_1 + \tfrac{1}{2}\theta_2,$$

$$E\left[\frac{X_{(r)}-\theta_1}{\theta_2}\right]=\frac{r}{n+1}-\frac{1}{2}.$$

7. Let $X_1, X_2, ..., X_n$ be a random sample with common p.d.f.

$$f(x)=\begin{cases}1, & 0<x<1\\ 0, & \text{otherwise}\end{cases}$$

(*i*) Find the p.d.f., mean and variance of $X_{(1)}$.
(*ii*) Find the p.d.f., mean and variance of $X_{(n)}$.
(*iii*) Find Corr. $(X_{(1)}, X_{(n)})$.

Ans. (*i*) $f_1(x)=n(1-x)^{n-1},\ 0\le x\le 1\,;\ E(X_{(1)})=1/(n+1)$
$\text{Var}(X_{(1)})=n/[(n+2)(n+1)^2]$

(*ii*) $f_n(x)=n x^{n-1}\,;\ 0\le x\le 1\,;\quad E(X_{(n)})=n/(n+1)$
$\text{Var}(X_{(n)})=n/[(n+2)(n+1)^2]$

(*iii*) **Hint.** $r(X_{(1)}, X_{(n)})=\dfrac{\text{Cov}(X_{(1)}, X_{(n)})}{\sqrt{\text{Var}(X_{(1)})\,.\,\text{Var}\,X_{(n)}}}$ [c.f. Chapter 10]

$$E(X_{(1)}\,.\,X_{(n)})=\int_0^1\int_0^y xy\,f_{1n}(x,y)\,dx\,dy=n(n-1)\int_0^1\int_0^y xy\,(y-x)^{n-2}\,dx\,dy$$

$$\left(\because f_{1n}(x,y)=n(n-1)f(x)\,[F(y)-F(x)]^{n-2}f(y)\,;\,0\le x<y\le 1\right)$$

$$\therefore\ E(X_{(1)}\,.\,X_{(n)})=n(n-1)\int_0^1\int_0^y x\,y^{n-1}\left(1-\frac{x}{y}\right)^{n-2}dx\,dy$$

$$=n(n-1)\int_0^1\int_0^1 y^{n+1}\,t\,(1-t)^{n-2}\,dt\,dy\,;\ \left(\frac{x}{y}=t\right)$$

$$=1/(n+2) \qquad \text{[On simplification]}$$

$\text{Cov}(X_{(1)}, X_{(n)})=1/[(n+1)^2\,.\,(n+2)]$
$\text{Corr.}(X_{(1)}, X_{(n)})=1/n$

8. Show that the c.d.f. of the mid-point (or mid-range) $M=\frac{1}{2}(X_{(1)}+X_{(n)})$, in a random sample of size n from a continuous population with c.d.f. $F(x)$ is:

$$F(m)=P(M\le m)=n\int_{-\infty}^{m}[F(2m-x)-F(x)]^{n-1}\,.\,f(x)\,dx$$

9. Let $X_i\ (i=1, 2, ..., n)$, be i.i.d. non-negative r.v.'s of continuous type. If $M_n=X_{(n)}=Max\,(X_1, X_2, ..., X_n)$, and $E(|X|)<\infty$, then prove that

$$E(M_n)=E(M_{n-1})+\int_0^\infty F^{n-1}(x)\,[1-F(x)]\,dx$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

Hence find $E(M_n)$ if X_i's are i.i.d. exponential variates with parameter λ.

Ans. $E(M_n) = \frac{1}{\lambda}\left[1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}\right]$

10. Show by means of an example that there may exist a r.v. X for which $E(X)$ does not exist but $E(X_{(r)})$ exists for some r;

Hint. Let $X_1, X_2, \ldots, X_n$ be a random sample from the popln. with p.d.f.

$$f(x) = \frac{1}{x^2},\ 1 < x < \infty;\quad F(x) = 1 - \frac{1}{x}$$

$E(X)$ does not exist, but $E(X_{(r)})$ exists for any $r < n$.

11. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with p.d.f.

$$f(x) = 1,\ 0 < x < 1$$
$$= 0,\ \text{otherwise}$$

Show that $Y_1 = X_{(1)}/X_{(2)},\ Y_2 = X_{(2)}/X_{(3)}, \ldots, \ldots,$

$Y_{n-1} = X_{(n-1)}/X_{(n)}$ and $Y_n = X_{(n)}$ are independently distributed and identify their distributions.

Hint. Proceed as in the hint to Q. No. 18 and 19.

12. Let X_1, X_2, X_3 be a random sample of size 3 from exponential distribution with parameter λ. Show that $Y_1 = X_{(3)} - X_{(2)}$ and $Y_2 = X_{(2)}$ are independently distributed.

Hint. $n = 3$; write joint p.d.f. of order statistics $X_{(2)}$ and $X_{(3)}$ and then transform to Y_1 and Y_2.

13. Let $X_1, X_2, \ldots, X_{2m+1}$ be an odd-size random sample from a $N(\mu, \sigma^2)$ population. Find the p.d.f. of the sample median and show that it is symmetric about μ, and hence has the mean μ.

14. A random sample of size n is drawn from an exponential population:

$$p(x) = \frac{1}{\theta} e^{-x/\theta};\ \theta > 0,\ x \ge 0$$

(*i*) Obtain the p.d.f. of $X_{(r)}$.

(*ii*) Show that $X_{(r)}$ and $W_{rs} = X_{(s)} - X_{(r)},\ r < s$ are independently distributed.

(*iii*) Identify the distribution of $W_1 = X_{(r+1)} - X_{(r)}$

[Gujrat Univ. M.Sc. (Stat.), 1991]

15. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a uniform population with p.d.f.

$$f(x) = \begin{cases} 1, & \text{if } 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Show that:

(*a*) $X_{(r)}$ is a $\beta_1(r, n-r+1)$ variate.

(*b*) $W_{rs} = X_{(s)} - X_{(r)}$ also has a Beta distribution which depends only on $s - r$ and not on s and r individually.

Ans. $f(w_{rs}) = \frac{1}{\beta(s-r, n-s+r+1)} \cdot w_{rs}^{s-r-1}(1-w_{rs})^{n-s+r};\ 0 \le w_{rs} \le 1$

16. In a random sample of size n from uniform $U[0, 1]$ population, obtain the p.d.f. of $W_{rs} = X_{(s)} - X_{(r)}$ and identify its distribution.

[Delhi Univ. M.Sc. (Stat.), 1987]

17. Obtain the p.d.f. of the range in a random sample of size 5 from the population with p.d.f. $e^{-x}, x > 0$. **[Meerut Univ. M.Sc. (Stat.), 1993]**

18. Let $X_1, X_2, ..., X_n$ be a random sample from continuous population with p.d.f.

$$f(x) = \beta e^{-x\beta};\ x \le 0,\ \beta > 0$$
$$= 0, \quad \text{otherwise}$$

(a) Show that $X_{(r)}$ and $X_{(s)} - X_{(r)}$ are independent for any $s > r$.

(b) Find the p.d.f. of $X_{(r+1)} - X_{(r)}$

(c) Let $Z_1 = n X_{(1)},\ Z_2 = (n-1)(X_{(2)} - X_{(1)}),$

$$Z_3 = (n-2)(X_{(3)} - X_{(2)}), ..., Z_n = (X_{(n)} - X_{(n-1)}) \qquad ...(*)$$

Show that $(Z_1, Z_2, ..., Z_n)$ and $(X_1, X_2, ..., X_n)$ are identically distributed.

Hint. (c) The joint p.d.f. of $X_{(1)}, X_{(2)}, ..., X_{(n)}$ is:

$$f(x_1, x_2, ..., x_n) = n!\, f(x_1) f(x_2) \dots f(x_n)$$
$$= n!\, \beta^n . e^{-\beta x_1} . e^{-\beta x_2} \dots e^{-\beta x_n} \qquad ...(**)$$

Transformation (*) gives:

$$X_{(1)} = \frac{Z_1}{n};\ X_{(2)} = \frac{Z_1}{n} + \frac{Z_2}{n-1};\ X_{(3)} = \frac{Z_1}{n} + \frac{Z_2}{n-1} + \frac{Z_3}{n-2}, \qquad ...(***)$$

$$X_{(n)} = \frac{Z_1}{n} + \frac{Z_2}{(n-1)} + \dots + \frac{Z_{n-1}}{2} + \frac{Z_n}{1}$$

$$J = \frac{\partial (X_{(1)}, X_{(2)}, ..., X_{(n)})}{\partial (Z_1, Z_2, ..., Z_n)} = \frac{1}{n!}$$

$0 < X_{(1)} < X_{(2)} < \dots < X_{(n)} < \infty \quad \Rightarrow \quad 0 < Z_i < \infty;\ i = 1, 2, ..., n.$

Using (***) and $|J|$, (**) transforms to

$$g(z_1, z_2, ..., z_n) = \left(\beta e^{-\beta Z_1}\right)\left(\beta e^{-\beta Z_2}\right) \dots \left(\beta e^{-\beta Z_n}\right);\ 0 < Z_i < \infty$$

$\Rightarrow \quad Z_1, Z_2, ..., Z_n$ are i.i.d. exponential variates with parameter β.

19. Let $X_1, X_2, ..., X_n$ be i.i.d. with p.d.f.

$$f(x) = \frac{1}{\sigma} \exp\left[-\left(\frac{x-\theta}{\sigma}\right)\right];\ x > 0$$
$$= 0, \quad \text{otherwise}$$

Show that $X_{(1)},\ X_{(2)} - X_{(1)}, X_{(3)} - X_{(2)}, ..., X_{(n)} - X_{(n-1)}$ are independent.

Hint. $\qquad Z_1 = X_{(1)};\ Z_2 = X_{(2)} - X_{(1)},\ ...,\ Z_n = X_{(n)} - X_{(n-1)}$

$\Rightarrow \qquad X_{(1)} = Z_1, X_{(2)} = Z_1 + Z_2, ..., X_{(n)} = Z_1 + Z_2 + \dots + Z_n$

$$J = \frac{\partial (X_{(1)}, X_{(2)}, ..., X_{(n)})}{\partial (Z_1, Z_2, ..., Z_n)} = 1$$

As in above problem

$$g(z_1, z_2, ..., z_n) = n! \prod_{i=1}^{n} f(x_i)\, |J|$$

$$= \left[\frac{n}{\sigma} e^{-n(z_1-\theta)/\sigma}\right] \left[\frac{(n-1)}{\sigma} e^{-(n-1)z_2/\sigma}\right] \cdots \left[\frac{2}{\sigma} e^{-2z_{n-1}/\sigma}\right] \times \left[\frac{1}{\sigma} e^{-z/\sigma}\right]$$

$\Rightarrow$ $Z_1, Z_2, ..., Z_n$ are independently distributed.

20. Find the p.d.f. of rth order statistic.

Let $X_1, X_2, ..., X_n$ be i.i.d. with a distribution function

$$F(y) = \begin{cases} y^{\alpha} \text{ if } 0 < y < 1 & ; \alpha > 0 \\ 0, \quad \text{otherwise} \end{cases}$$

Show that $\dfrac{X_{(i)}}{X_{(n)}}$, $i = 1, 2, ..., n-1$ and $X_{(n)}$ are independent.

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

21. Let $x_{i1}, x_{i2}, ..., x_{in}, i = 1, 2, ..., k$ be k random samples from $R\left(\frac{1}{2}, 1\right)$ population. Find the distribution of $U = x_{1(n)} \cdot x_{2(n)} ...... x_{k(n)}$, where $x_{i(n)}$ is the maximum of ith sample.

(R = Rectangular population)

[Delhi Univ. M.Sc. (Stat.), 1989]

22. For the exponential distribution $f(x) = e^{-x}$, $x \geq 0$, find the p.d.f. of the range W in a random sample of size n and show that

$$E(W) = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1}$$

Ans. $g(w) = (n-1) e^{-w} (1 - e^{-w})^{n-2}$; $w \geq 0$.

23. Let $X_1, X_2, ..., X_n$ be a random sample of size n from $U[a, b]$ population. Obtain the p.d.f's of (*i*) $X_{(1)}$, (*ii*) $X_{(n)}$ and (*iii*) joint p.d.f. of $X_{(1)}$ and $X_{(n)}$.

24. Let X_1, X_2 be i.i.d. r.v.'s with p.d.f.

$$P(X_i = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, ...; \; i = 1, 2$$

where $\lambda > 0$. Let $M = \text{Max.}(X_1, X_2)$ and $N = \text{Min}(X_1, X_2)$

Find the marginal p.m.f.'s of M and N.

8·15. Truncated Distributions. Let X be a random variable with p.d.f. (or p.m.f.) $f(x)$. The distribution of X is said to be truncated at the point $X = a$ if all the values of $X \leq a$ are discarded. Hence the p.d.f. (or p.m.f.) $g(\,.\,)$ of the distribution, truncated at $X = a$ is given by:

$$g(x) = \frac{f(x)}{P(X > a)}; \quad x > a \qquad ...(8·71)$$

$$= \frac{f(x)}{\sum\limits_{x > a} f(x)}; \quad x > a \text{ (For discrete r.v.)} \qquad ...(8·71\,a)$$

$$= \frac{f(x)}{\int_a^{\infty} f(x)\,dx}; \quad x > a \text{ (For continuous r.v.)} \qquad \ldots(8\cdot 71\ b)$$

For the continuous r.v. X, the rth moment (about origin) for the truncated distribution is given by:

$$\mu_r' = E(X^r) = \int_a^{\infty} x^r g(x)\,dx = \frac{\int_a^{\infty} x^r f(x)\,dx}{\int_a^{\infty} f(x)\,dx} \qquad \ldots(8\cdot 72)$$

Example 8·54 *Let $X \sim B(n, p)$. Find the mean and variance of the binomial distribution truncated at $X = 0$.*

Solution. Let $f(x)$ be the p.m.f. of $X \sim B(n, p)$ variate. Then the p.m.f. $g(x)$ of the Binomial distribution truncated at $X = 0$ is given by:

$$g(x) = \frac{f(x)}{P(X > 0)} = \frac{f(x)}{1 - P(X = 0)} = \frac{f(x)}{1 - f(0)}$$

$$= \frac{1}{1 - q^n} \cdot {}^nC_x p^x q^{n-x}; \quad x = 1, 2, \ldots, n \qquad \ldots(*)$$

$$E(X) = \sum_{x=1}^{n} x\, g(x) = \frac{1}{1 - q^n} \sum_{x=1}^{n} x \cdot {}^nC_x p^x q^{n-x}$$

$$= \frac{1}{1 - q^n} \left[\sum_{x=0}^{n} x \cdot {}^nC_x p^x q^{n-x} - 0 \right]$$

$$= np/(1 - q^n) \qquad \ldots(**)$$

$$E(X^2) = \frac{1}{1 - q^n} \sum_{x=1}^{n} x^2 \cdot {}^nC_x p^x q^{n-x}$$

$$= \frac{1}{1 - q^n} \left[\sum_{x=0}^{n} x^2 \cdot {}^nC_x p^x q^{n-x} - 0 \right]$$

$$= \frac{1}{1 - q^n} \left[npq + n^2 p^2 \right]$$

$$\left[\because X \sim B(n, p);\ E(X^2) = \text{Var } X + (EX)^2 = npq + n^2 p^2 \right]$$

$$\therefore \quad \text{Var}(X) = EX^2 - [E(X)]^2$$

$$= \frac{1}{1 - q^n} \left[npq + n^2 p^2 - \frac{n^2 p^2}{1 - q^n} \right]$$

Example 8·55 *Obtain the mean and variance of a standard Cauchy distribution truncated at both ends, with relevant range of variation as $(-\beta, \beta)$.*

[Delhi Univ. M.Sc. (Stat.), 1987]

Solution. Let $f(x)$ be the p.d.f. of standard Cauchy distribution. Then the p.d.f. $g(x)$ of the truncated distribution with relevant range of variation as $(-\beta, \beta)$ is given by:

$$g(x) = \frac{f(x)}{P(-\beta \le X \le \beta)} = \frac{f(x)}{\int_{-\beta}^{\beta} f(x)\, dx}$$

$$= \frac{\frac{1}{\pi} \frac{1}{1+x^2}}{\frac{1}{\pi}\int_{-\beta}^{\beta} \frac{dx}{1+x^2}} = \frac{1}{1+x^2} \cdot \frac{1}{\left| \tan^{-1} x \right|_{-\beta}^{\beta}}$$

$$= \frac{1}{2\tan^{-1}\beta} \cdot \frac{1}{(1+x^2)}\,; \quad -\beta \le x \le \beta \qquad \ldots(*)$$

$$\text{Mean} = \int_{-\beta}^{\beta} x\, g(x) = \frac{1}{2\tan^{-1}\beta} \int_{-\beta}^{\beta} \frac{x}{1+x^2}\, dx$$

$$= 0 \qquad [\because \text{ Integrand is an odd function of } x]$$

$$\text{Variance} = \mu_2' - \mu_1'^2 = \mu_2' = \int_{-\beta}^{\beta} x^2 g(x)\, dx$$

$$= \frac{2}{2\tan^{-1}\beta} \int_0^{\beta} \frac{x^2}{1+x^2}\, dx = \frac{1}{\tan^{-1}\beta} \int_0^{\beta} \left(1 - \frac{1}{1+x^2}\right) dx$$

$$= \frac{1}{\tan^{-1}\beta} \left| x - \tan^{-1} x \right|_0^{\beta} = \frac{1}{\tan^{-1}\beta} \left(\beta - \tan^{-1}\beta\right)$$

$$= \frac{\beta}{\tan^{-1}\beta} - 1$$

Example 8·56 *Consider a truncated standard normal distribution truncated at both ends with relevant range of variation as (A, B). Obtain the p.d.f., mkean, mode and variance of the truncated distribution.*

[Delhi Univ. M.Sc. (Stat.), 1988]

Solution. Let $Z \sim N(0, 1)$ with p.d.f. $\varphi(z)$ and c.d.f. $\Phi(z) = P(Z \le z)$. Then the p.d.f. $g(\,.\,)$ of the truncated normal distribution is given by:

$$g(z) = \frac{\varphi(z)}{P(A \le Z \le B)} = \frac{\varphi(z)}{\Phi(B) - \Phi(A)} = \frac{1}{k}\varphi(z)\,; \; A \le z \le B \qquad \ldots(i)$$

where $\quad k = \Phi(B) - \Phi(A)$.

$$\text{Mean} = \int_A^B z\, g(z)\, dz = \frac{1}{k} \int_A^B z\, \varphi(z)\, dz \qquad \ldots(ii)$$

We have $\quad \varphi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-z^2/2}$

$$\Rightarrow \qquad \frac{d}{dz}\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,(-z) = -z\,\varphi(z)$$

$$\Rightarrow \quad \int z\,\varphi(z)\,dz = -\varphi(z) \qquad \ldots(*)$$

Substituting in (ii)

$$\text{Mean} = \frac{1}{k}\left| -\varphi(z) \right|_A^B = \frac{\varphi(A) - \varphi(B)}{\Phi(B) - \Phi(A)} = \mu' \text{ (say)} \qquad \ldots(iii)$$

$$\text{Mode} = \begin{cases} B & \text{if } A < 0,\ B < 0 \\ 0 & \text{if } A < 0,\ B > 0 \\ A & \text{if } A > 0,\ B > 0 \end{cases}$$

$$\text{Variance} = \int_A^B z^2 g(z)\, dz - \mu'^2$$

$$= \frac{1}{k}\left[\left| z(-\varphi(z)) \right|_A^B + \int_A^B \varphi(z)\, dz \right] - \mu'^2$$

$$= \frac{-1}{k}\left[B\,\varphi(B) - A\,\varphi(A) - \{\Phi(B) - \Phi(A)\} \right] - \mu'^2$$

$$= 1 + \frac{A\,\varphi(A) - B\,\varphi(B)}{\Phi(B) - \Phi(A)} - \mu'^2$$

where μ' is given in (*iii*).

EXERCISE 8 (k)

1. Find the mean and variance of the truncated Poisson distribution with parameter λ, truncated at the origin.

Ans. p.d.f.

$$g(x) = \frac{1}{1 - e^{-\lambda}}\left[\frac{e^{-\lambda}\lambda^x}{x!} \right],\ x = 1, 2, 3, \ldots;\ E(X) = \frac{\lambda}{1 - e^{-\lambda}};\ EX^2 = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}}$$

2. Obtain the p.d.f. and the mean of truncated standard normal distribution, for positive values only.

Ans. $g(z) = 2 \cdot \dfrac{1}{\sqrt{2\pi}}\, e^{-z^2/2} = 2\,\varphi(z)\,;\; z > 0\,;\; E(Z) = \sqrt{2/\pi}$

3. (*a*) Let X be normally distributed with mean μ and variance σ^2. Truncate the density of X on the left at a and on the right at b, and then calculate the mean of the truncated distribution. [Note that if $a = \mu - c$ and $b = \mu + c$, then the mean of the truncated distribution should equal μ.]

[Delhi Univ. B.Sc. (Maths. Hons.), 1989]

(*b*) If X is normally distributedwith mean μ and variance σ^2, find the mean of the conditional distribution of X given $a \le X \le b$.

Hint. In fact Problem in Part (*a*) is same as in Part (*b*), stated differently.

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-\mu)^2/2\sigma^2}\ ;\ -\infty < x < \infty$$

Mean of truncated distribution is

$$\mu' = \frac{\int_a^b x f(x)}{\int_a^b f(x)\,dx} = \frac{\int_a^b (x - \mu + \mu) f(x)\,dx}{\int_a^b f(x)\,dx}$$

$$= \mu + \frac{\int_a^b (x-\mu)\, f(x)\,dx}{\int_a^b f(x)\,dx} = \mu + \sigma^2 \left[\frac{f(a) - f(b)}{F(b) - F(a)}\right]$$

where $F(x) = P(X \le x)$, is the distribution function of X.

$$\left\{ \because \quad f'(x) = f(x) \times \left[-\left(\frac{x-\mu}{\sigma^2}\right)\right] \quad \Rightarrow \quad -\sigma^2 f'(x) = (x-\mu) f(x) \right.$$

$$\left. \Rightarrow \quad \int_a^b (x-\mu) f(x)\,dx = -\sigma^2 \,\Big|\, f(x) \,\Big|_a^b = \sigma^2 [f(a) - f(b)] \right\}$$

4. A truncated Poisson distribution is given by the mass function

$$f(x) = \frac{1}{1 - e^{-\lambda}} \cdot \frac{e^{-\lambda}\lambda^x}{x!}\ ;\ x = 1, 2, 3, \ldots,\ \lambda > 0$$

Find the m.g.f. and hence mean and variance of the distribution.

5. Consider the p.d.f.

$$g(x) = \frac{f(x)}{1-F(x_0)},\ x > x_0 \qquad \ldots(*)$$

where $f(x) = (\sqrt{2\pi}\,.\,\sigma)^{-1}\,.\exp\left[-(x-\mu)^2/2\sigma^2\right], -\infty < \mu < \infty, \sigma > 0$

and $F(x_0) = \int_{-\infty}^{x_0} f(u)\,du$

[(*) is the p.d.f. of $N(\mu, \sigma^2)$ distribution, truncated at the point $x = x_0$].
Show that the first two raw moments can be expressed as:

$$\mu_1' = \mu + \lambda\sigma\ ;\ \mu_2' = \mu^2 + \lambda\sigma(x_0 + \mu) + \sigma^2$$

where $\lambda\left[1 - F(x_0)\right] = f\left[(x_0 - \mu)/\sigma\right]$

6. Let $X \sim \gamma(\alpha)$. Obtain the p.d.f. of the truncated distribution, truncated at the point x_0 and prove that $\mu_r' = EX^r$ for the truncated gamma distribution

$$= \left[EX^r \text{ for the untruncated } \gamma(\alpha) \text{ distribution}\right] \times \left[\left\{1 - I_{x_0}(\alpha+\gamma)\right\}/\left\{1 - I_{x_0}(\alpha)\right\}\right]$$

where $I_{x_0}(\alpha) = \int_0^{x_0} \frac{e^{-x}x^{\alpha-1}}{\Gamma\alpha}\,dx$ (Incomplete Gamma Integral)

Ans. $g(x) = \frac{1}{1 - I_{x_0}(\alpha)}\cdot\left(\frac{e^{-x}\,.\,x^{\alpha-1}}{\Gamma\alpha}\right);\ x > x_0$

7. Explain the concept of 'Truncation'. For a standard normal distribution, truncated at both ends with relevant range of variation as $[A, B]$, obtain mean, variance and mean deviation about mean. **[Delhi Univ. M.Sc. (Stat.), 1983]**

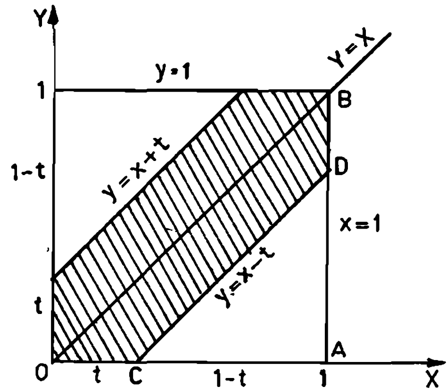
ADDITIONAL EXERCISES ON CHAPTER VIII

1. If the random variable X has the density function $f(x) = x/2$, $0 < x < 2$, find the r^{th} moment of X^2. Deduce that $Z = X^2$ has the distribution $g(y) = 1/4$, $0 < y < 4$.

2. Suppose that X is a random variable for which $E(X) = \mu$ and $V(X) = \sigma^2$. Further suppose that Y is uniformly distributed over the interval (α, β). Determine α and β so that $E(X) = E(Y)$ and $V(X) = V(Y)$.

3. A boy and a girl agree to meet at a certain park between 4 and 5 P.M. They agree that the one arriving first will wait t hours, $0 \le t \le 1$, for the other to arrive. Assuming that the arrival times are independent and uniformly distributed, find the probability that they meet.

Hence obtain the probability that they meet if $t = 10$ minutes.



Hint. $p = P(|X - Y| \le t) = \dfrac{\text{Shaded Area}}{\text{Total Area}}$

$= 2\,[\text{ Area } OAB - \text{Area } CAD\,]/1 \times 1$

$= 1 - (1-t)^2 = 2t - t^2$

Ans. $t = 10$ minutes, Probability $= 11/36$

4. Let $X \sim U[0, 1]$. Find Corr. (X, Y), where $Y = X^n$

[Delhi Univ. B.A. (Hons.) (Spl. Course-Statistics), 1989]

5. (*a*) Let $X_1, X_2, \ldots, X_n$ be independent random variables having a common rectangular distribution over the interval $[a, b]$. Obtain the distribution of

$$Y = \max.\ (X_1, X_2, \ldots, X_n)$$

(*b*) Let $X \sim U\{0, 1, 2, \ldots, r\}$, $r = ab$, $a > 1$, $b < r$, where a and b are positive integers. Show that the distribution of X coincides with $U + V$, where U and V are independent r.v.'s both with uniform distribution on appropriate subsets of $\{0, 1, \ldots, ab\}$. **(Indian Civil Services, 1988)**

6. If $X_1, X_2, \ldots, X_n$ are mutually independent rectangular variates on $[0,1]$, prove that the density function of $X_1 . X_2 . \ldots . X_n$ is

$$f(x) = \begin{cases} \dfrac{(-\log x)^{n-1}}{(n-1)!}, & 0 < x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

7. (*a*) Let X_1 and X_2 be independent r.v.'s, each uniform on $[0,1]$. Show that:

$Y_1 = \sqrt{-2 \log X_1}\ .\ (\sin 2\pi X_2)$ and $Y_2 = \sqrt{-2 \log X_1}\ .\ \cos(2\pi X_2)$ are independent r.v.'s, and that each is $N(0, 1)$.

[This is known as Box and Muller transformation .]

(*b*) Given a sequence of independent r.v.'s $X_1, X_2, \ldots$ which are uniform on $[0,1]$, produce a sequence of independent r.v.'s $Y_1, Y_2, \ldots$ that are $N(0, 1)$ and independent.

(You may assume that the sum of two independent Normal distributions is itself Normally distributed.)

8. (*a*) Assume a random variable X has a standard normal distribution and let $Y = X^2$

(*i*) Show that $F_Y(t) = \dfrac{2}{\sqrt{2\pi}} \displaystyle\int_0^{\sqrt{t}} e^{-u^2/2}\, du,\ t \geq 0$

(*ii*) Determine $F_Y(t)$ when $t < 0$ and describe the density function $f_Y(t)$.

(*b*) Let $\Phi(x)$ be the standard normal distribution function and let

$$\Phi(x) = \int_{-\infty}^{x} \varphi(u)\, du$$

Show that

(*i*) $\left(\dfrac{1}{x} - \dfrac{1}{x^2}\right) \varphi(x) \leq 1 - \Phi(x) \leq \dfrac{1}{x} \varphi(x),\ x > 0$

(*ii*) $\lim\limits_{x \to \infty} \dfrac{x[1 - \Phi(x)]}{\varphi(x)} = 1$

9. (*a*) If X_1 and X_2 are independent normal variates with means μ_1 and μ_2 and variance σ_1^2 and σ_2^2, respectively, find the relation between α, β, γ and δ so that

$P(c_1 X_1 + c_2 X_2 < \alpha) = \gamma$ and $P(c_1 X_1 + c_2 X_2 < \beta) = \delta$

(*b*) If X and Y are independent normal variates with equal means and standard deviations 9 and 12 respectively, and if

$$P[X + 2Y < 3] = P[2X - Y \geq 4],$$

determine the common mean of X and Y.

10. If $X \sim N(0, 1)$ and A is constant, obtain the characteristic function of $(X - A)^2$, Hence or otherwise prove that

$$\kappa_r = 2^{r-1}(1 + rA^2) \cdot (r-1)!$$

where κ_r is the rth cumulant of X.

Ans. $\Phi_{(X-A)^2}(t) = (1 - 2it)^{-1/2} \cdot \exp[\, it\, a^2/(1 - 2it)\,]$

11. (*a*) If X is a standard normal variate and α is a small number, prove that $P(X \leq \alpha) = \frac{1}{2} + \dfrac{8}{\sqrt{2\pi}}\left(\alpha - \dfrac{\alpha^3}{2^3 \cdot 3} + \dfrac{1}{2!} \cdot \dfrac{\alpha^5}{2^5 \cdot 5} - \dots\right)$

(*b*) Show that

$$\int_0^x e^{-x^2}\, dx = xe^{-x^2}\left[1 + \frac{1}{3}(2x^2) + \frac{1}{3.5}(2x^2)^2 + \dots\right]$$

12. Let X be log-normal variate with p.d.f.

$$f(x, \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-(\log x - \mu)^2/2\sigma^2\right\},\ x > 0, \sigma > 0, |\mu| < \infty$$

Show that $e^a X^b$ has a lognormal p.d.f. $f(x, a + b\mu, b\sigma)$. If $X_1, X_2, ..., X_n$ are n independent observations on X, then show that $G = (X_1 X_2 \ldots X_n)^{1/n}$ also has a log-normal p.d.f. $f(x, \mu, \sigma/n)$.

13. (*a*) The distribution

$$dF = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), -\infty < x < \infty$$

is transformed by the transformation $X = a \log_e (Y - b) + C$. Find the distribution of Y. Evaluate the mean, the mode and the median of this distribution of (Y) and arrange them in order of magnitude when $b > 0$.

(*b*) If $Y = a \log (X - b) + C$ has normal distribution with mean zero and unit variance, obtain the distribution of X and evaluate its mean, median and mode.

14. A standard variable X is transformed to Y by the relation

$$X = \frac{1}{c} [\log_{10} Y - a]$$

with $m = e^{b^2 c^2}$ and $\frac{1}{b} = \log_{10} e$; show that for the transformed variable

$$\beta_1 = m^2 (m + 3) - 4 \text{ and } \beta_2 = m^2 (m^2 + 2m + 3) - 3$$

15. If X and Y are independent normal variates with zero expectations and variances σ_1^2 and σ_2^2, show that $Z = XY/\sqrt{X^2 + Y^2}$ is normal with variance $\sigma_z^2 = \left[(1/\sigma_1^2) + (1/\sigma_2^2)\right]^{-1}$

16. If X_i ; $i = 1, 2, ..., n$ is a random sample of size n from a normal population with mean μ and variance σ^2, obtain the joint distribution of

$$U = \sum_{i=1}^{n} a_i X_i \text{ and } V = \sum_{i=1}^{n} b_i X_i$$

where a_i's and b_i's are arbitrary constants.

Hence or otherwise show that the necessary and sufficient condition that U and V are independent is that $\Sigma\, a_i b_i = 0$.

17. Let X_1, X_2 and X_3 be three independent normal variates with the same mean μ and variance σ^2. Let

$$Y_1 = \frac{X_1 - X_2}{\sqrt{2}}, \; Y_2 = \frac{X_1 - 2X_2 + X_3}{\sqrt{6}}, \text{ and } Y_3 = \frac{X_1 + X_2 + X_3}{\sqrt{3}}$$

Show that Y_1, Y_2 and Y_3 are independent normal variates.

Show that

$$Y_1^2 + Y_2^2 = \sum_{i=1}^{3} (X_i - \bar{X})^2, \text{ where } \bar{X} = \frac{X_1 + X_2 + X_3}{3}.$$

18. A random variable X has probability density function

$$f(x) = C \cdot \varphi(x), x \geq k$$

where k is a given number, C is a constant chosen to ensure that $f(x)$ is a probability density function and

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\tfrac{1}{2}x^2\right\}$$

If $g(k) = \int_k^{\infty} \varphi(x)\,dx$, show that the arithmetic mean and the variance of X are respectively $\dfrac{\varphi(k)}{g(k)}$ and $1 + \dfrac{\varphi(k)}{g(k)}\left\{k - \dfrac{\varphi(k)}{g(k)}\right\}$

19. Three independent observations X_1, X_2, X_3 are given from a univariate $N(m, \sigma^2)$. Derive the joint sampling distribution of :

(a) $U = X_1 - X_3$; (b) $V = X_2 - X_3$

(c) $W = X_1 + X_2 + X_3 - 3m$

Deduce the p.d.f. of $Z = U/V$. Show that mode $Z = 1/2$ and obtain the significance of this modal value. (Indian Civil Services, 1986)

20. *Neyman's Contagious (Compound) Distribution.* Let $X \sim P(\lambda\, y)$ where y itself is an observation of a variate $Y \sim P(\lambda_1)$. Find the unconditional distribution of X and show that its mean is less than its variance.

21. If $X_1, X_2, \ldots, X_n$ are independent random variables, having the probability law $p(x_i) = \dfrac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i\,!}$; $i = 1, 2, \ldots, n, x_i = 0, 1, 2, \ldots, \infty$

and if $X = \sum_{i=1}^{n} X_i$ and $\lambda = \sum_{i=1}^{n} \lambda_i$

then under certain conditions to be specified clearly,

$$P\left\{\frac{X - \lambda}{\sqrt{\lambda}} < \alpha\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-t^2/2}\,dt \text{ as } n \to \infty .$$

22. Prove that

$$\frac{1}{n\,!}\int_{\lambda}^{\infty} e^{-x} x^n\,dx = \sum_{r=0}^{n} \frac{e^{-\lambda}\lambda^r}{r\,!} ; n = 0, 1, 2, \ldots$$

Using the above, write down the relation connecting the distribution functions of a Poisson and a Gamma variate.

23. A two-dimensional random variable (X, Y) has the joint p.d.f.,

$$f(x, y) = \frac{1}{\Gamma(\mu)\,\Gamma(\nu)}\, x^{\mu - 1} (y - x)^{\nu - 1} e^{-y}$$

in the $x - y$ plane where $0 < x < y < \infty$ and zero elsewhere. Show that the marginal distributions of X and Y are gamma distributions.

24. Starting from a suitable urn model, deduce the differential equation of the Pearsonian curves in the form

$$\frac{1}{y}\cdot\frac{dy}{dx}=\frac{a+x}{b_0+b_1x+b_2x^2}$$

Also discuss the limitations of ranges in the solution of such differential equations.

25. Karl Pearson showed that the differential equation

$$\frac{d\,[f(x)]}{f(x)}=\frac{d-x}{a+bx+cx^2}\,dx$$

yields most of the important frequency curves when appropriate values of $a, b, c,$ and d are chosen. Show that

(*i*) when $d=0$ and $a=c=0$ as well as $b>0$, the differential equation yields exponential distribution,

(*ii*) when $b=c=0$ and $a>0$, the differential equation yields normal distribution, and

(*iii*) when $a=c=0$, $b>0$ and $d>-b$, the differential equation yields gamma distribution.

26. Find the m.g.f. of the distribution with p.d.f.

$$f(x)=\left(\frac{\lambda}{2\pi x^3}\right)^{1/2}\exp\left[\frac{-\lambda(x-\mu)^2}{2\mu^2x}\right], x>0, y>0, \mu>0 \qquad \ldots(*)$$

Also show that rth cumulant is given by:

$$K_r=1.3.5\ldots(2r-3)\,\mu^{2r-3}\lambda^{1-r}$$

(*) is the p.d.f. of *Standard Inverse Gaussian distribution.*

27. Obtain $M_X(t)$, when $X\sim P(\lambda)$. Find the limit as $\lambda\to\infty$ of m.g.f. of $(X-\lambda)/\sqrt{\lambda}$ and interpret the result in the context of C.L.T. Also prove that:

$$\lim_{\lambda\to\infty}\sum_{k=\alpha}^{\beta}\left[\frac{e^{-\lambda}.\lambda^k}{k!}\right]=\frac{1}{\sqrt{2\pi}}\int_a^b\exp\left(-\frac{1}{2}u^2\right)du,$$

$$\alpha=\lambda+a\sqrt{\lambda},\quad \beta=\lambda+b\sqrt{\lambda}$$

Show that $2X$ is not a Poisson variate. Give a set of conditions under which $X+Y$ too is a Poisson variate. (Indian Civil Services, 1983)

28. The random variables X_k $(k=1, 2, \ldots,)$ are independent and have the Cauchy distribution with p.d.f.

$$f(x)=\frac{1}{\pi}\cdot\frac{1}{1+x^2}, -\infty<x<\infty. \text{ Let } Y_n=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

Examine whether the sequence $\{Y_n\}$ obeys the weak law of large numbers.

29. Let $X_1, X_2, \ldots, X_n, \ldots$ be independent Bernoulli variates such that

$P(X_k=1)=p_k,\quad P(X_k=0)=q_k=(1-p_k),\ k=1, 2, \ldots, n, \ldots$

Examine whether the sequence $\left\{ \sum_{i=1}^{n} X_i \right\}$ follows the central limit theorem.

OBJECTIVE TYPE QUESTIONS

1. Choose the correct answer from B and match it with each item in A.

A	B
(a) β_2 for a Normal distribution	(1) $3\sigma^4$
(b) β_1 for a Normal distribution	(2) 0
(c) μ_3 for a Normal distribution	(3) 3
(d) μ_4 for a Normal distribution	(4) $e^{i\mu t - t^2\sigma^2/2}$
(e) Characteristic function for Normal distribution	(5) $\frac{4}{5}\sigma$
(f) Moment generating function of Normal distribution	(6) 0
(g) Mode of Normal distribution	(7) $e^{\mu t + t^2\sigma^2/2}$
(h) Mean deviation from mean for Normal distribution	(8) μ

II. Match the distributions:

(a) Uniform distribution (1) $f(x) = \dfrac{1}{\lambda\left\{(x-\mu)^2 + \lambda^2\right\}}, -\infty < x < \infty$

(b) Normal distribution (2) $f(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty$

(c) Exponential distribution (3) $f(x) = \dfrac{1}{b-a}, a \le x \le b$

(d) Beta distribution (4) $f(x) = \dfrac{1}{B(m, n)} x^{m-1}(1-x)^{n-1}, 0 \le x \le 1$

(e) Cauchy distribution (5) $f(x) = \dfrac{1}{a} e^{-x/a}, x \ge 0$

III. If X_i $(i = 1, 2, 3, \ldots, n)$ are independent $N(0, 1)$, write (without proof), the distribution of

(i) $X_1 - 2X_2 + X_3$, (ii) $\dfrac{X_2}{X_3}$, (iii) $\sum_{i=1}^{n} X_i$ (iv) $\dfrac{X_1^2}{X_1^2 + X_2^2}$,

(v) $\dfrac{X_1^2}{X_2^2 + X_3^2}$, (iv) $\dfrac{X_i}{X_j}, (i \ne j)$ (vii) $\sum_{i=1}^{n} X_i^2$ (viii) $\dfrac{X_i^2}{X_j^2}, i \ne j$

IV. In each case, specify the distribution for which:

(i) Moments do not exist.

(ii) Mean = variance.

(*iii*) Mean < variance

(*iv*) Mean > variance.

(*v*) $\phi(t) = e^{it - t^2}$

(*vi*) $\phi(t) = e^{-|t|}$

V. State the conditions under which

(*i*) Binomial distribution,

(*ii*) Poisson distribution

tends to Normal distribution.

VI. (*i*) Give two examples of variates which you expect to be distributed normally.

(*ii*) Give two examples of variates which you expect to be distributed exponentially.

VII. State which of the following statements are TRUE and which are FALSE. In case of false statements, give the correct statement.

(*i*) For normal distribution mean deviation about mean is greater than quartile deviation.

(*ii*) X is a random variable following Cauchy distribution, for which mean does not exist but variance exists.

(*iii*) In case of normal distribution $\beta_1 = 3$, $\beta_2 = 0$,

(*iv*) If X and Y are two independent normal variates, then $X - Y$ is also a normal variate.

(*v*) Binomial distribution tends to normal distribution as $n \to \infty$.

(*vi*) For normal distribution, mean = mode = median.

(*vii*) It is possible to reduce every normal distribution to the standard normal distribution by a transformation.

(*viii*) In uniform distribution, the percentile points are equi-spaced.

(*ix*) Normal distribution is symmetrical only for some specified values of the mean and variance.

(*x*) Normal distribution can be obtained as a limiting case of Poisson distribution with the parameter $\lambda \to \infty$.

VIII. Give the correct answer to each of the following :

(*i*) The mean and variance of Normal distribution

(*a*) are same, (*b*) cannot be same, (*c*) are sometimes equal, (*d*) are equal in the limiting case, as $n \to \infty$.

(*ii*) The mean and variance of Gamma distribution

(*a*) are same, (*b*) cannot be same, (*c*) are sometimes equal, (*d*) are equal in the limiting case, as $n \to \infty$.

(*iii*) X is normally distributed with zero mean and unit variance. The variance of X^2 is

(*a*) 0, (*b*) 1, (*c*) 2, (*d*) 4

(*iv*) The points of inflexion of Normal curve are

(*a*) $m \pm \sigma$, (*b*) $m \pm 2\sigma$ (*c*) $m \pm 3\sigma$ (*d*) $m \pm \frac{2}{3}\sigma$

(v) The moment generating function of gamma distribution is
(a) $(1+t)^{\lambda}$, (b) $(1-t)^{\lambda}$, (c) $(1-t)^{-\lambda}$, (d) $(1+t)^{-\lambda}$

(vi) The characteristic function of Cauchy distribution is
(a) e^{-t}, (b) $e^{-|t|}$, (c) e^{t}, (d) $e^{|t|}$

(vii) Area to the right of the point x_1 is 0·6 and to the left of the point x_2, is 0·7. Which is the correct:
(i) $x_1 > x_2$, (ii) $x_1 < x_2$ or (iii) $x_1 = x_2$?

(viii) The standard normal distribution is represented by
(a) $N(0, 0)$, (b) $N(1, 1)$, (c) $N(0, 1)$, (d) $N(1, 0)$.

(ix) For a normal distribution, quartile deviation, mean deviation, standard deviation are in the ratio
(a) $\frac{4}{5}:\frac{2}{3}:1$, (b) $\frac{2}{3}:\frac{4}{5}:1$, (c) $1:\frac{4}{5}:\frac{2}{3}$, (d) $\frac{2}{3}:1:\frac{4}{5}$

(x) The normal distribution is a limiting form of binomial distribution if
(a) $n \to \infty$, $p \to 0$, (b) $n \to 0, p \to q$, (c) $n \to \infty, p \to n$, (d) $n \to \infty$ and neither p nor q is small.

(xi) Normal curve is
(a) very flat, (b) bell shaped symmetrical about mean, (c) very peaked, (d) smooth.

(xii) The normal distribution is a limiting case of Poisson's when
(a) $\lambda \to 0$, (b) $\lambda \to \sigma$, (c) $\lambda \to \infty$, (d) $\lambda < \sigma$.

(xiii) In a normal curve the number of observations less than mean are included in the range
(a) $\bar{x} \pm 3\,\sigma$, (b) $\bar{x} \pm 1{\cdot}96$, (c) $\bar{x} \pm 2\,\sigma$, (d) $\bar{x} \pm 0{\cdot}67\,\sigma$

(xiv) If X is a standard normal variate, then $\frac{1}{2}X^2$ is a
(a) Gamma variate with parameter $\frac{1}{2}$, (b) a normal variate, (c) a Poisson variate.

(xv) The range of the beta variate is
(a) $(0, \infty)$, (b) $(-\infty, \infty)$, (c) $(0, 1)$, (d) $(-1, +1)$

IX. Fill in the blanks :—

(i) The mean deviation of normal distribution is...
(ii) The p.d.f. of Gamma distribution is...
(iii) The relationship between Beta distributions of first and second kind is...
(iv) The normal distribution is a limiting form of binomial distribution if...
(v) Mean = variance for ... distribution (continuous).
(vi) For the normal distribution :

$$\beta_1 = \ldots$$
$$\beta_2 = \ldots$$
$$\text{Mean deviation} = \ldots$$

Quartile deviation = ...

(*vii*) The characteristic function of a Gamma distribution is ...

(*viii*) The points of inflexion for a normal curve are ...

(*ix*) For the normal distribution with varaince σ^2,

$$\mu_{2r} = \ldots$$

$$\mu_{2r+1} = \ldots$$

(*x*) A normal distribution is completely specified by the parameters...

(*xi*) For normal distribution

S.D. : M.D. : Q.D. : : ... : ... : ...

(*xii*) If $X \sim N(\mu, \sigma^2)$ then

$$P(\mu - \sigma < X < \mu + \sigma) = \ldots$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \ldots$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \ldots$$

(*xiii*) If X is a random variable with distribution function F then $F(X)$ has ... distribution.

(*xiv*) If $X \sim N(0, 1)$, then $X^2/2$ has ... distribution with parameter ...

X. Random variables X_i are independent and all of them have the same distribution defined by

$$f(x) = \frac{1}{\sqrt{8\pi}} \exp\left\{-(x-1)^2/8\right\}, \ -\infty < x < \infty$$

Find the distribution of

(*i*) $\sum_{i=1}^{10} X_i/10$ and (*ii*) $X_1 - 2X_2 + X_3$

Ans. (*i*) $N\left(1, \frac{2}{5}\right)$, (*ii*) $N(0, 24)$

XI. The random variables X_i, $i = 1, 2, \ldots$ are independent and all of them have the same distribution defined by

$$f(x) = \frac{1}{\sqrt{18\pi}} e^{-(x-1)^2/18}; -\infty < x < \infty.$$

Find the distribution of

(*a*) $\frac{1}{5} \sum_{i=1}^{5} X_i$.

(*b*) $3X_1 - X_2 + 2X_3$.

XII. Find the mean and standard deviation of a probability distribution whose frequency function is given by

$$f(x) = Ce^{-(1/24)(x^2 - 6x + 9)}, -\infty < x < \infty$$

where C is a cosntant. **[Delhi Univ. B.A. (Stat. Hons.), 1986]**

Ans. Mean 3. $\sigma^2 = 12$, $C = \frac{1}{\sqrt{24\pi}}$

XIII. (a) If $f(x) = ke^{-3x^2+6x}$, $-\infty < x < \infty$,

obtain the values of k, μ and σ^2.

Ans. $k = \sqrt{(3/\pi)}\, e^{-3}$; $\mu = 1, \sigma^2 = \frac{1}{6}$

(b) If X is anormal variate with mean μ and variance σ^2, find the distribution of $Y = aX + b$.

Ans. $Y \sim N(a\mu + b, a^2\sigma^2)$.

(c) If X is distributed normally with mean μ and standard deviation σ, write down the distribution of $U = 2X - 3$ and find the mean and variance of U.

Ans. $U \sim N(2\mu - 3, 4\sigma^2)$

XIV. If X_1 is normally distributed with a mean 10 and variance 16 and X_2 is nromally distributed with a mean 10 and variance 15 and if $W = X_1, + X_2$, what will be the values of the two parameters of the distribution of the variate W? (Assume that X_1 and X_2 are independent).

(c) Let X and Y be independent and normally distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Find the mean and variance of $\frac{1}{2}(X + Y)$

XV. Write a note on the role of Normal distribution in Statistics.

XVI. If X_i, $(i = 1, 2, 3, ..., n)$ are *i.i.d.* standard Cauchy variates, write the distribution of $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Ans. Standard Cauchy.

XVII. Is the sum of two independent Cauchy variates a Cauchy varaite?

If X_i. $(i = 1\ 2, 3, 4)$ are independent standard normal variates, what is the distribution of $\frac{X_1}{X_2} + \frac{X_3}{X_4}$?

Ans. Cauchy.

XVIII. "The role of Cauchy distribution in statistical theory often lies in providing counter examples." Elucidate.

XIX. Write a note on the role of Central Limit Theorem in Statistics.

XX. If X_i, $(i = 1, 2, 3, 4)$ are *i.i.d.* $N(0, 1)$, write the distribution of:

(i) $X_1 - X_2$ (ii) $X_1 + X_2$ (iii) $\frac{X_1}{X_2}$ (iv) $\frac{X_1}{X_2} + \frac{X_3}{X_4}$

(v) $X_1 - X_2 + X_3 - X_4$ (vi) $\frac{1}{2}X_1^2$ (vii) $\frac{1}{2}(X_2^2 + X_3^2)$

(viii) $\frac{X_1^2}{X_1^2 + X_2^2}$ (ix) $\frac{X_3^2}{X_4^2}$ (x) $\frac{X_1^2}{X_2^2 + X_3^2}$.

Ans. (i) $N(0, 2)$; (ii) $N(0, 2)$; (iii) Standard Cauchy; (iv) Cauchy; (v) $N(0, 4)$; (vi) $\gamma(\frac{1}{2})$; (vii) $\gamma(1)$; (viii) $\beta_1(\frac{1}{2}, \frac{1}{2})$; (ix) $\beta_2(\frac{1}{2}, \frac{1}{2})$; (x) $\beta_2(\frac{1}{2}, 1)$.

CHAPTER NINE

Curve Fitting and Principle of Least Squares

9·1. Curve Fitting. Let (x_i, y_i) ; $i = 1, 2, ..., n$ be a given set of n pairs of values, X being independent variable and Y the dependent variable. The general problem in curve fitting is to find, if possible, an analytic expression of the form $y = f(x)$, for the functional relationship suggested by the given data.

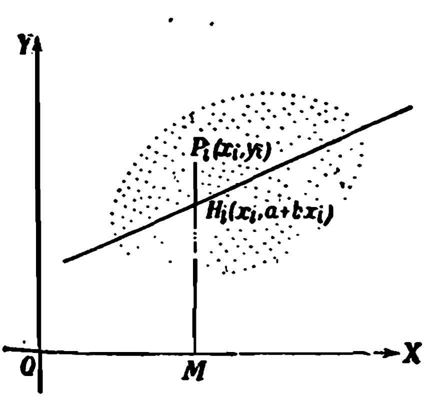
Fitting of curves to a set of numerical data is of considerable importance– theoretical as well as practical. Theoretically it is useful in the study of correlation and regression, *e.g.*, lines of regression can be regarded as fitting of linear curves to the given bivariate distribution (c.f. § 10·8·1). In practical statistics it enables us to represent the relationship between two variables by simple algebraic expressions, *e.g.*, polynomials, exponential or logarithmic functions. Moreover, it may be used to estimate the values of one variable which would correspond to the specified values of the other variable.

9·1·1. Fitting of a straight line. Let us consider the fitting of a straight line

$$Y = a + bX \qquad ...(9·1)$$

to a set of n points (x_i, y_i) ; $i = 1, 2, ..., n$. Equation (9·1) represents a family of straight lines for different values of the arbitrary constants 'a' and 'b'. The problem is to determine 'a' and 'b' so that the line (9·1) is the line of "*best fit*".

The term '*best fit*' is interpreted in accordance with Legender's principle of least squares which consists in minimising the sum of the squares of the deviations of the actual values of y from their estimated values as given by the line of best fit.



Let $P_i(x_i, y_i)$ be any general point in the scatter diagram (§ 10·2). Draw $P_i M \perp$ to x–axis meeting the line, (9·1) in H_i. Abscissa of H_i is x_i and since H_i lies on (9·1), its ordinate is $a + bx_i$. Hence the co-ordinates of H_i are $(x_i, a + bx_i)$.

$$P_i H_i = P_i M - H_i M$$
$$= y_i - (a + bx_i),$$

is called the *error of estimate* or the *residual* for y_i.

According to the principle of

least squares, we have to determine a and b so that

$$E = \sum_{i=1}^{n} P_i H_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

is minimum. From the principle of maxima and minima, the partial derivatives of E, with respect to (w.r.t.) a and b should vanish separately, *i.e.*,

$$\frac{\partial E}{\partial a} = 0 = -2 \sum_{i=1}^{n} (y_i - a - bx_i) \text{ and } \frac{\partial E}{\partial b} = 0 = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) \quad \text{...(9·2)}$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \text{ and } \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \quad \text{...(9·2a)}$$

Equations (9·2) and (9·2*a*) are known as the *normal equations* for estimating a and b.

All the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i y_i$, can be obtained from the given set of points (x, y_i); $i = 1, 2, ..., n$ and the equations (9·2*a*) can be solved for a and b. With the values of a and b so obtained, equation (9·1) is the line of best fit to the given set of points (x_i, y_i); $i = 1, 2, ..., n$.

Remark. The equation of the line of best fit of y on x is obtained on Climinating a and b in (9·1) and (9·2*a*) and can be expressed in the determinant form as follows:

$$\begin{vmatrix} Y & X & 1 \\ \Sigma y_i & \Sigma x_i & n \\ \Sigma x_i y_i & \Sigma x_i^2 & \Sigma x_i \end{vmatrix} = 0 \quad \text{...(9·2b)}$$

9·1·2. Fitting of second degree parabola. Let

$$Y = a + bX + cX^2 \quad \text{...(9·3)}$$

be the second degree parabola of best fit to set of n points (x_i, y_i); $i = 1, 2, ..., n$. Using the principle of least squares, we have to determine the constants a, b and c so that

$$E = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

is minimum.

Equating to zero the partial derivatives of E with respect to a, b and c separately, we get the normal equations for estimating a, b and c as

$$\left.\begin{aligned} \frac{\partial E}{\partial a} &= 0 = -2 \Sigma (y_i - a - bx_i - cx_i^2) \\ \frac{\partial E}{\partial b} &= 0 = -2 \Sigma x_i (y_i - a - bx_i - cx_i^2) \\ \frac{\partial E}{\partial c} &= 0 = -2 \Sigma x_i^2 (y_i - a - bx_i - cx_i^2) \end{aligned}\right\} \quad \text{...(9·4)}$$

$$\Rightarrow \quad \begin{aligned} \Sigma y_i &= na + b\,\Sigma x_i + c\,\Sigma x_i^2 \\ \Sigma x_i y_i &= a\,\Sigma x_i + b\,\Sigma x_i^2 + c\,\Sigma x_i^3 \\ \Sigma x_i^2 y_i &= a\,\Sigma x_i^2 + b\,\Sigma x_i^3 + c\,\Sigma x_i^4, \end{aligned} \qquad \text{...(9·4}a\text{)}$$

summation taken over i from 1 to n.

For given set of points (x_i, y_i); $i = 1, 2, ..., n$, equations (9·4a) can be solved for a, b and c, and with these values of a, b and c, (9·3) is the parabola of best fit.

Remark. Eliminating a, b and c in (9·3) and (9·4a), the parabola of best fit of y on x is given by

$$\begin{vmatrix} Y & X^2 & X & 1 \\ \Sigma y_i & \Sigma x_i^2 & \Sigma x_i & n \\ \Sigma x_i y_i & \Sigma x_i^3 & \Sigma x_i^2 & \Sigma x_i \\ \Sigma x_i^2 y_i & \Sigma x_i^4 & \Sigma x_i^3 & \Sigma x_i^2 \end{vmatrix} = 0 \qquad \text{...(9·4}b\text{)}$$

9·1·3. Fitting of Polynomial of kth Degree. If

$$Y = a_0 + a_1 X + a_2 X^2 + ... + a_k X^k \qquad \text{...(9·5)}$$

is the k^{th} degree polynomial of best fit to the set of points (x_i, y_i); $i = 1, 2, ..., n$, the constants $a_0, a_1, a_2, ..., a_k$ are to be obtained so that

$$E = \sum_{1=j}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2 - ... - a_k\, x_i^k)^2$$

is minimum. Thus the normal equations for estimating $a_0, a_1, ..., a_k$ are obtained on equating to zero the partial derivatives of E w.r.t. $a_0, a_1, ..., a_k$ separately, *i.e.*,

$$\left.\begin{aligned} \frac{\partial E}{\partial a_0} &= 0 = -2\,\Sigma(y_i - a_0 - a_1\, x_i - a_2\, x_i^2 - ... - a_k x_i^{\,k}) \\ \frac{\partial E}{\partial a_1} &= 0 = -2\,\Sigma\, x_i (y_i - a_0 - a_1\, x_i - a_2\, x_i^2 - ... - a_k x_i^{\,k}) \\ \frac{\partial E}{\partial a_k} &= 0 = -2\,\Sigma\, x_i^k (y_i - a_0 - a_1\, x_i - a_2\, x_i^2 - ... - a_k x_i^{\,k}) \end{aligned}\right\} \qquad \text{...(9·6)}$$

$\Rightarrow$

$$\left.\begin{aligned} \Sigma\, y_i &= na_0 + a_1\, \Sigma x_i + a_2\, \Sigma x_i^2 + \cdots + ak\, \Sigma x_i^k \\ \Sigma\, x_i y_i &= a_0\, \Sigma x_i + a_1\, \Sigma x_i^2 + a_2\, \Sigma x_i^3 + \cdots + ak\, \Sigma x_i^{k+1} \\ \Sigma\, x_i^k y_i &= a_0\, \Sigma x_i^k + a_1\, \Sigma x_i^{k+1} + a_2\, \Sigma x_i^{k+2} + \cdots + ak\, \Sigma x_i^{2k}, \end{aligned}\right\} \qquad \text{...(9·6}a\text{)}$$

summation extended over i from 1 to n. These are $(k + 1)$ equations in $(k + 1)$ unknowns $a_0, a_1, a_2, ..., a_k$ and can be solved with the help of algebra.

Remark. It has been found that in all the above cases, the values of the second order derivatives, *viz.*, $\dfrac{\partial^2 E}{\partial a_0^2}, \dfrac{\partial^2 E}{\partial a_1^2} \cdots$ come out to be positive at the points $a_0, a_1, ..., a_k$, the solutions of the 'normal equations'. Hence they provide minima of E. For proof see Remark 1 to § 10.7.1–Lines of Regression.

Example 9·1. *Fit a straight line to the following data.*

X :	1	2	3	4	6	8
Y :	2·4	3	3·6	4	5	6

Solution. Let the line be $Y = a + bX$

	X	Y	X^2	XY
	1	2·4	1	2·4
	2	3·0	4	6·0
	3	3·6	9	10·8
	4	4·0	16	16·0
	6	5·0	36	30·0
	8	6·0	64	48·0
Total	24	24	130	113·2

Using normal equations (9.2a), we get

$$24 = 6a + 24b \text{ and } 113{\cdot}2 = 24a + 130b$$

Solving these equations, we get $a = 1{\cdot}976$ and $b = 0{\cdot}506$.

Example 9·2. *Fit a parabola of second degree to the following data* :

X :	0	1	2	3	4
Y :	1	1·8	1·3	2·5	6·3

(Delhi Univ. B.Sc., Oct. 1992)

Solution. Let $Y = a + bX + cX^2$ be the second degree parabola.

	X	Y	X^2	X^3	X^4	XY	X^2Y
	0	1·0	0	0	0	0	0
	1	1·8	1	1	1	1·8	1·8
	2	1·3	4	8	16	2·6	5·2
	3	2·5	9	27	81	7·5	22·5
	4	6·3	16	64	256	25·2	100·8
Total	10	12·9	30	100	354	37·1	130·3

Using normal equations (9·4), we get

$12{\cdot}9 = 5a + 10b + 30c$; $37{\cdot}1 = 10a + 30b + 100c$;

$130{\cdot}3 = 30a + 100b + 354c$

Solving these equations, we get $a = 1{\cdot}42$, $b = -\ 1{\cdot}07$ and c 0·55. Thus the required equation of the second degree parabola is

$$Y = 1{\cdot}42 - 1{\cdot}07\,X + 0{\cdot}55X^2$$

Remark. If the values which X and Y take are large, the calculation of $\Sigma\, x$, $\Sigma\, x^2$, $\Sigma\, x\, y$, ..., becomes quite tedious and the solution of the normal equations, is also quite cumbersome. In this case arithmetic is reduced to a great extent by suitable change of origin in X or (and) in Y.

9·1·4. Change of origin. Let us suppose that the values of X are given to be equidistant at an interval of h, *i.e.*, X takes the values, (say), $a, a+h, a+2h, \ldots$ If n is odd, *i.e.*, $n = 2m+1$ (say), we take

$$U = \frac{X - (\text{ middle term })}{\text{Interval}} = \frac{X-(a+mh)}{h}$$

Now U takes the values $-m, -(m-1), \ldots, -1, 0, 1, \ldots, (m-1), m,$ so that $\Sigma U = \Sigma U^3 = 0$.

If n is even, *i.e.*, $n = 2m$ (say), then there are two middle terms, *viz.*, mth and $(m+1)$th terms which are $a+(m-1)h$ and $a+mh$. In this case, we take

$$U = \frac{X - (\text{mean of two middle terms})}{\frac{1}{2}(\text{interval})} = \frac{X - [a + \frac{1}{2}(2m-1)h]}{\frac{1}{2}(h)}$$

$$= \frac{2X - 2a - (2m-1)h}{h} \qquad \ldots(9{\cdot}7)$$

Now for $X = a, a+h, \ldots, a+(2m-1)h$; U takes the values $-(2m-1), -(2m-3), \ldots, -3, -1, 1, 3, \ldots, (2m-3), (2m-1)$.

Again we see that $\Sigma U = \Sigma U^3 = 0$.

Example 9·3. *The weights of a calf taken at weekly intervals are given below. Fit a straight line using the method of least squares and calculate the average rate of growth per week.*

Age (X) :	1	2	3	4	5	6	7	8	9	10
Weight (Y) :	52·5	58·7	65·0	70·2	75·4	81·1	87·2	95·5	102·2	108·4

Solution. Let the variables age and weight be denoted by X and Y respectively.

Here $n = 10$, *i.e.*, even and the values of X are equidistant at an interval of unity, *i.e.*, $h = 1$. Thus we take

$$U = \frac{X - \{(5+6)/2\}}{\frac{1}{2}} = 2X - 11.$$

Let the least-square line of Y on U be $Y = a + bU$.

The normal equations for estimating a and b are

$$\Sigma Y = na + b\Sigma U \quad \text{and} \quad \Sigma UY = a\Sigma U + b\Sigma U^2$$

X	Y	U	U^2	UY
1	52·5	−9	81	−472·5
2	58·7	−7	49	−410·9
3	65·0	−5	25	−325·0
4	70·2	−3	9	−210·6
5	75·4	−1	1	−75·4
6	81·1	1	1	81·1
7	87·2	3	9	261·6

8	95·5	5	25	477·5
9	102·2	7	49	715·4
10	108·4	9	81	975·6
Total	796·2	0	330	1016·8

Thus the normal equations are

$$796{\cdot}2 = 10a + 0 \times b \quad \text{and} \quad 1016{\cdot}8 = a \times 0 + 330b,$$

which give $\quad a = 79{\cdot}62 \quad$ and $\quad b = \dfrac{1016{\cdot}8}{330} = 3{\cdot}08$ (approx).

$\therefore$ The least square line of Y on U is

$$Y = 79{\cdot}62 + 3{\cdot}08U$$

Hence the line of best fit of Y on X is

$$Y = 79{\cdot}62 + 3{\cdot}08\,(2X - 11) \quad \Rightarrow \quad Y = 45{\cdot}74 + 6{\cdot}16\,X$$

The weights of the calf (as given by the line of best fit $Y = A + BX$) after 1, 2, 3, ... weeks are $A + B, A + 2B, A + 3B, \ldots,$ respectively. Hence the average rate of growth per week is B units, *i.e.*, 6·16 units.

EXERCISE 9 (a)

1. *(a)* (x_i, y_i) ; $i = 1, 2, \ldots, n$, give the co-ordinates of n points in a plane. It is proposed to fit a straight line $Y = aX + b$ to those points such that the sum of the squares of the perpendiculars from those n points to the line is a minimum. Find the constants a and b. Use the above method to fit a straight line to the following points :

X :	0	1	2	3	4
Y :	1	1·8	3·3	4·5	6·3

Ans. $\quad Y = 0{\cdot}72 + 1{\cdot}33X$

(b) Fit a straight line of the form $Y = AX + B$ to the following data:

X :	0	5	10	15	20	25	30
Y :	10	14	19	25	31	36	39

2. Show that the line of best fit to the following data is given by

$$Y = -0{\cdot}5X + 8$$

X :	6	7	7	8	8	8	9	9	10
Y :	5	5	4	5	4	3	4	3	3

3. *(a)* How do you define the term "line of best fit". Give the normal equations generally used to obtain such a line. Fit a straight line and parabolic curve to the following data :X :1·01·52·02·53·03·54·0

Y :	1·1	1·3	1·6	2·6	2·7	3·4	4·1

Ans. $\quad Y = 1{\cdot}04 - 0{\cdot}20X + 0{\cdot}24X^2$

(b) Fit a straight line to the following data. Plot the observed and the ex-

(b) Fit a straight line to the following data. Plot the observed and the expected values in a graph and examine whether the straight line gives an adequate fit

x...	1	2	3	4	5	6	7	8
y...	55	46	40	38	33	30	29	30

4. An experiment is conducted to verify the law of falling under gravity expressed by
$$S = \frac{1}{2} g t^2$$
where S is the distance fallen at time t and g is a gravitational constant. The following results are obtained :

t (seconds) :	1	2	3	4	5
S (feet) :	15	70	140	250	380

Taking S as the depenent variable, fit a straight line to the data by the method of least squares in a manner that you can estimate g. What is the estimate of g ?

5. *(a)* Explain the method of fitting a second degree parabola by using the principle of least squares.

(b) Fit a parabola $Y = a + bx + cx^2$ to the following data :

X ;	1	2	3	4	5	6	7
Y :	2·3	5·2	9·7	16·5	29·4	35·5	54·4

6. Fit a second degree parabola to the following data taking X as the independent variable :

X :	1	2	3	4	5	6	7	8	9
Y :	2	6	7	8	10	11	11	10	9

Ans. $Y = -1 + 3{\cdot}55X - 0{\cdot}27X^2$

7. In a spectroscopic method for determining the per cent X of natural rubber-content of vulcanizates, the variable Y used is $1 + \log_{10} r$, where r is the ratio of transmission at two selected wavelengths. In order to establish a relationship between X and Y, the following data were obtained :

X :	0	20	40	60	80	100
Y :	2·19	2·65	3·16	3·57	3·93	4·27

Using least square method, fit a parabola. Comment on your results.

8. Fit a second degree curve $Y = a + bX + cX^2$ to the following data relating to profit of a certain company.

Year :	1980	1982	1984	1986	1988
Profit in lakhs of rupees :	125	140	165	195	230

Estimate the profit in the year 1995.

Ans. $Y = 114 + 7{\cdot}2X + 3{\cdot}15X^2$

9. Explain the method of least squares of fitting a curve to the given mass of data :

X :	−2	−1	0	1	2

Y : y_1 y_2 y_3 y_4 y_5

Fit a parabola $Y = a + bX + c\,(X^2 - 2)$, by the method of least squares and show that

$$a = \bar{y},\quad b = \frac{1}{10}(-2y_1 - y_2 + y_4 + 2y_5),\quad c = \frac{1}{14}(2y_1 - y_2 + y_4 + 2y_5)$$

10. Show that the best fitting linear function for the points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) may be expressed in the form

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix} = 0,\ (i = 1, 2, .., n)\,.$$

Show that the line passes through the mean point $(\bar{x}, \bar{y})$.

9·2. Most Plausible Solution of a System of Linear Equations. Method of least squares is helpful in finding the most plausible values of the variables satisfying a system of independent linear equations whose number is more than the number of variables under study. Consider the following set of m equations in n variables $X, Y, Z, ..., T$:

$$\left.\begin{aligned} a_1X + b_1Y + c_1Z + \dots + k_1T &= l_1 \\ a_2X + b_2Y + c_2Z + \dots + k_2T &= l_2 \\ &\vdots \\ a_mX + b_mY + c_mZ + \dots + k_mT &= l_m \end{aligned}\right\} \quad ...(9{\cdot}8)$$

where $a_i, b_i, ..., l_i$; $i = 1, 2, ..., m$ are constants.

If $m = n$, the system of equations (9·8) can be solved uniquely with the help of algebra. If $m > n$, it is not posible to determine a unique solution $X, Y, Z, ..., T$ which will satisfy the system (9·8). In this case we find the values of $X, Y, Z, ..., T$ which will satisfy the system (9·8) as nearly as possible.

Legender's principle of least squares consists in minimising the sum of the squares of the 'residuals' or the 'errors'. If

$$E_i = a_iX + b_iY + c_iZ + \dots + k_iT - l_i\ ;\ i = 1, 2, .., m$$

is the residual for the ith equation, then we have to determine $X, Y, Z, ..., T$ so that

$$U = \sum_{i=1}^{m} E_i^2 = \sum_{i=1}^{m} (a_i X + b_iY + c_i Z + \dots + k_i T - l_i)^2$$

is minimum.

Using the principle of maxima and minima in differential calculus, the partial derivatives of 'U' w.r.t. $X, Y, Z, ..., T$ should vanish separately. Thus

$$\left.\begin{aligned}\frac{\partial U}{\partial X}&=0=\sum_{i=1}^{m} a_i\,(a_i X+b_iY+c_iZ+\ldots+k_iT-l_i)\\ \frac{\partial U}{\partial Y}&=0=\sum_{i=1}^{m} b_i\,(a_i X+b_iY+c_iZ+\ldots+k_iT-l_i)\\ &\vdots\qquad\qquad \vdots\\ \frac{\partial U}{\partial T}&=0=\sum_{i=1}^{m} k_i\,(a_i X+b_iY+c_iZ+\ldots+k_iT-l_i)\end{aligned}\right\}\quad\ldots(9\cdot9)$$

These are known as the *normal equations* for $X, Y, Z, \ldots, T$ respectively. Thus we have n – *normal equations in* n unknowns $X, Y, Z, \ldots, T$ and their unique solution gives the best or the most plausible solution of the system (9·8).

Here we see that *the normal equation for any variable is obtained by multiplying each equation by the coefficient of the variable in that equation and then adding all the resulting equations.*

Example 9·4. *Find the most plausible values of X and Y from the following equations :*

$$X-5Y+4=0,\qquad 2X-3Y+5=0$$
$$X+2Y-3=0,\qquad 4X+3Y+1=0$$

Solution. Normal equation for X is

$$1\,.\,(X-5Y+4)+2\,(2X-3Y+5)+1\,.\,(X+2Y-3)+4\,(4X+3Y+1)=0$$

$$\Rightarrow\qquad 22X+3Y+15=0\qquad\ldots(*)$$

Normal equation for Y is

$$-5\,(X-5Y+4)-3\,(2X-3Y+5)+2\,(X+2Y-3)+3\,(4X+3Y+1)=0$$

$$\Rightarrow\qquad 3X+47Y-38=0\qquad\ldots(**)$$

Solving (*) and (**), we get $X=-0{\cdot}799$ and $Y=0{\cdot}86$.

Hence the most plausible values of X and Y are $X=-0{\cdot}80$ (approx.) and $Y=0{\cdot}86$ (approx.)

EXERCISE 9 (b)

1. Find the most plausible values of X and Y from the following equations:

(i) $X+Y=3{\cdot}01,\quad 2X-Y=0{\cdot}03,\quad X+3Y=7{\cdot}03,\quad 3X+Y=4{\cdot}97.$

Ans. $X=1{\cdot}0003,\ Y=2{\cdot}0007.$

(ii) $X+Y=3,\quad X-Y=2,\quad X+2Y-4=0,\quad X=2Y+1.$

2. Find the most plausible values of X, Y and Z from the following equations :

$X-Y+2Z=3,\ 3X+2Y-5Z=5,\ 4X+Y+4Z=21$ and $-X+3Y+3Z=14$

Ans. $X=2{\cdot}47,\ Y=3{\cdot}55,\ Z=1{\cdot}92.$

9·3. Conversion of Data to Linear Form. Sometimes it may happen that the original data is not in a linear form but can be reduced to linear form by some simple transformation of variables. We will illustrate this by considering the following curves :

(a) Fitting of a Power Curve. $Y = aX^b$ ...(9·10)

to a set of n points.

Taking logarithm of both sides, we get

$$\log Y = \log a + b \log X$$

$$\Rightarrow \qquad U = A + bV$$

where $U = \log Y$, $A = \log a$ and $V = \log X$.

This is a linear equation in V and U.

Normal equations for estimating A and B are

$$\Sigma U = nA + b\Sigma V \quad \text{and} \quad \Sigma UV = A\Sigma V + b\Sigma V^2 \qquad ...(9\cdot10a)$$

These equations can be solved for A and b and consequently, we get

$$a = \text{antilog } (A)$$

With the values of a and b so obtained, (9·10) is the curve of best fit to the set of n points.

(b) Fitting of Exponential Curves. (i) $Y = ab^X$, (ii) $Y = ae^{bX}$ to a set of n points.

(i) $$Y = ab^X \qquad ...(9\cdot11)$$

Taking logarithm of both sides, we get

$$\log Y = \log a + X \log b$$

$$\Rightarrow \qquad U = A + BX$$

where $U = \log Y$, $A = \log a$ and $B = \log b$.

This is linear equation in X and U.

The normal equations for estimating A and B are

$$\Sigma U = nA + B\Sigma X \quad \text{and} \quad \Sigma XU = A\Sigma X + B\Sigma X^2 \qquad ...(9\cdot11a)$$

Solving these equations for A and B, we finally get

$$a = \text{antilog } (A) \quad \text{and} \quad B = \text{antilog } (B)$$

With these values of a and b, (9·11) is the curve of best fit to the given set of n points.

(ii) $$Y = ae^{bX} \qquad ...(9\cdot12)$$

$$\log Y = \log a + bX \log e = \log a + (b \log e)\, X$$

$$\Rightarrow \qquad U = A + BX$$

where $U = \log Y$, $A = \log a$ and $B = b \log e$.

This is linear equation in X and U.

Thus the normal equations are

$$\Sigma U = nA + B\Sigma X \quad \text{and} \quad \Sigma XU = A\Sigma X + B\Sigma X^2 \qquad ...(9\cdot12a)$$

From these we find A and B and consequently

$$a = \text{antilog } (A) \quad \text{and} \quad b = \frac{B}{\log e}$$

Example 9·5. *Fit an exponential curve of the form $Y = ab^x$ to the following data :*

X :	1	2	3	4	5	6	7	8
Y :	1·0	1·2	1·8	2·5	3·6	4·7	6·6	9·1

Solution.

	X	Y	$U = \log Y$	XU	X^2
	1	1·0	0·0000	0·0000	1
	2	1·2	0·0792	0·1584	4
	3	1·8	0·2553	0·7659	9
	4	2·5	0·3979	1·5916	16
	5	3·6	0·5563	2·7815	25
	6	4·7	0·6721	4·0326	36
	7	6·6	0·8195	5·7365	49
	8	9·1	0·9590	7·6720	64
Total	36	30·5	3·7393	22·7385	204

(9·11a) gives the normal equations as

$$3{\cdot}7393 = 8A + 36B$$

and $$22{\cdot}7385 = 36A + 204B$$

Solving, we get

$B = 0{\cdot}1408$ and $A = -0{\cdot}1662 = \bar{1}{\cdot}8338$

$\therefore\ b = \text{Antilog } B = 1{\cdot}383$ and $a = \text{Antilog } A = 0{\cdot}6821$

Hence the equation of the required curve is

$$Y = 0{\cdot}6821\,(1{\cdot}38)^X$$

Example 9·6. *Derive the least square equations for fitting a curve of the type $Y = aX + (b/X)$, to a set of n points (x_i, y_i) ; $i = 1, 2, \ldots, n$.*

Solution. The error of estimate E_i for the ith point (x_i, y_i) is given by

$$E_i = \left(y_i - ax_i - \frac{b}{x_i} \right)$$

According to the principle of least squares, we have to determine the values of a and b so that sum of the squares of errors E, *viz.*,

$$E = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \left(y_i - ax_i - \frac{b}{x_i} \right)^2$$

is minimum.

Consequently, the normal equations are

$$\frac{\partial E}{\partial a} = 0 = -2 \sum_{i=1}^{n} x_i \left(y_i - ax_i - \frac{b}{x_i} \right)$$

$$\frac{\partial E}{\partial b} = 0 = -2 \sum_{i=1}^{n} \frac{1}{x_i}\left(y_i - ax_i - \frac{b}{x_i} \right)$$

which on simplification give

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i^2 + nb$$

and $$\sum_{i=1}^{n} \left(\frac{y_i}{x_i} \right) = na + b \sum_{i=1}^{n} \left(\frac{1}{x_i^2} \right)$$

Example 9·7. *Three independent measurements on each of the three angles A, B, C of a triangle are as follows :*

A	B	C
39·5	60·3	80·1
39·3	62·2	80·3
39·6	60·1	80·4

Obtain the best estimates of the three angles taking into account the relation that the sum of the angles is equal to 180°.

Solution. Let the three observations on A be denoted by x_1, x_2, x_3, on B by y_1, y_2, y_3 and on C by z_1, z_2, z_3. Let θ_1, θ_2 be the best estimates for A and B respectively.

According to the principle of least squares, our problem is to estimate θ_1 and θ_2, so that

$$E = \Sigma (x_i - \theta_1)^2 + \Sigma (y_i - \theta_2)^2 + \Sigma (z_i - 180 + \theta_1 + \theta_2)^2$$

is minimum, summation being taken over i from 1 to 3.

Equating to zero the partial derivatives of E w.r.t. θ_1 and θ_2, the normal equations are

$$\frac{\partial E}{\partial \theta_1} = 0 = -\Sigma (x_i - \theta_1) + \Sigma (z_i - 180 + \theta_1 + \theta_2) \qquad \text{...}(*)$$

$$\frac{\partial E}{\partial \theta_2} = 0 = -\Sigma (y_i - \theta_2) + \Sigma (z_i - 180 + \theta_1 + \theta_2) \qquad \text{...}(**)$$

From (*) and (**), we get

$$\left.\begin{aligned} 3\theta_1 - \Sigma x_i + \Sigma z_i - 540 + 3\theta_1 + 3\theta_2 = 0 \\ 3\theta_2 - \Sigma y_i + \Sigma z_i - 540 + 3\theta_1 + 3\theta_2 = 0 \end{aligned}\right\} \qquad \text{...}(***)$$

But $\Sigma x_i = 39{\cdot}5 + 39{\cdot}3 + 39{\cdot}6 = 118{\cdot}4$

$\Sigma y_i = 60{\cdot}3 + 62{\cdot}2 + 60{\cdot}1 = 182{\cdot}6$

$\Sigma z_i = 80{\cdot}1 + 80{\cdot}3 + 80{\cdot}4 = 240{\cdot}8$

Substituting in (***), we get

$$6\theta_1 + 3\theta_2 - 417{\cdot}6 = 0 \quad \text{and} \quad 3\theta_1 + 6\theta_2 - 481{\cdot}8 = 0$$

$$\therefore \quad \hat{A} = \hat{\theta}_1 = 39{\cdot}27, \hat{B} = \hat{\theta}_2 = 60{\cdot}66 \text{ and } \hat{C} = 180 - \hat{\theta}_1 - \hat{\theta}_2 = 80{\cdot}07$$

9·4. Selection of Type of Curve to be Fitted. The greatest limitation of the method of curve fitting by the principle of least squares is the choice of the mathematical curve to be fitted to the given data. The choice of a particular curve for describing the given data requires great skill, intelligence and expertise. The graph of the given data enables us to have a fairly good idea about the type of the curve to be fitted. The graph will clearly reveal if the trend is linear (straight line) or curvilinear (non-linear). If the graph exhibits a curvilinear trend then further approximations to the type of trend curve can be obtained on plotting the data on a semi-logarithmic scale. A careful study of the graph obtained on plotting the data on an arithmetic or semi-logarithmic scale often provides adequate basis for selecting the type of the curve. The various types of curves that may be used to describe the given data in practise are: [If y_x is the value of the dependent variable corresponding to the value x of the independent variable]

(*i*) *A straight line:* $y_x = a + bx$

(*ii*) *Second degree parabola:* $y_x = a + bx + cx^2$

(*iii*) *kth degree polynomial :* $y_x = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k$

(*iv*) *Exponential curve:* $y_x = ab^x$

$$\Rightarrow \quad \log y_x = \log a + x \log b = A + Bx, \text{ (say).}$$

(*iv*) *Second degree curve fitted to logarithms:*

$$y_x = ab^x \, c^{x^2}$$

$$\Rightarrow \quad \log y_x = \log a + x \log b = x^2 \log c$$
$$= A + Bx + Cx^2, \text{ (say).}$$

(*vi*) *Growth Curves:*

(*a*) $y_x = a + b^x$ (Modified Exponential Curve)

(*b*) $y_x = abc^x$ (Gompertz Curve)

$$\Rightarrow \quad \log y_x = \log a + c^x \,.\, \log b = A + Bc^x, \text{ (say)}$$

(*c*) $y_x = \dfrac{k}{1 + \exp(a + bx)}$, $b < 0$ (Logistic Curve)

For decideing about the type of curve to be fitted to a given set of data, the following points may be helpful:

(*i*) When the y_x series is found to be increasing by equal absolute amounts, the straight line curve is used. In this case, the graph of the data will give a straight line graph.

(*ii*) The logarithmic straight line (exponential curve $y_x = ab^x$) is used when the series is increasing or decreasing by a constant percentage rather than a constant absolute amount. In this case, the data plotted on a semi-logarithmic scale will give a straight line graph.

(*iii*) Second degree curve fitted to logarithms may be tried if the data plotted on a semi-logarithmic scale is not a straight line graph but shows curvature, being concave either upward or downward.

For further guidelines, the following statistical tests based on the calculus of finite differences [*c.f.* Chapter 17] may be applied.

We know that for a polynomial y_x of nth degree in x,

$$\left.\begin{aligned}\Delta^r y_x &= \text{constant}, & r = n\\ &= 0 & , r > n\end{aligned}\right\}$$

where Δ is the difference operator given by $\Delta y_x = y_{x+h} - y_x$, h being the interval of differencing and $\Delta^r y_x$ is the rth order difference of y_x.

1. If $\Delta y_x =$ constant, use straight line curve.
2. If $\Delta^2 y_x =$ constant, use a second degree (parabolic) curve.
3. If $\Delta (\log y_x) =$ constant, use exponential curve.
4. If $\Delta^2 (\log y_x) =$ constant, use second degree curve fitted to logarithms.
5. If Δy_x tends to decrease by a constant percentage, use modified exponential curve.
6. If Δy_x resembles a skewed frequency curve, use a Gompertz curve or Logistic curve.
7. The growth curves, *viz.*, modified exponential, Gompertz and Logistic curves, can be approximated by the constancy of the ratios

$$\frac{\Delta y_x}{\Delta y_{x-1}}, \left\{\frac{\Delta \log y_x}{\Delta \log y_{x-1}}\right\}, \left\{\frac{\Delta (1/y_x)}{\Delta (1/y_{x-1})}\right\},$$

respectively for all possible values of x.

EXERCISE 9 (c)

1. Describe the method of fitting the following curves :

(*i*) $Y = ae^{bx}$, (*ii*) $Y = aX^b$

2. (*a*) Fit an equation of the form $Y = ab^X$ to the following data :

X :	2	3	4	5	6
Y :	144	172·8	207·4	248·8	298·6

Ans. $Y = (101{\cdot}3)\,(1{\cdot}196)^X$

(*b*) Fit a curve of the type $Y = ab^X$ to the following data :

X :	2	3	4	5	6
Y :	8·3	15·4	33·1	65·2	127·4

Estimate Y when $X = 4{\cdot}5, 7$ and $3{\cdot}5$.

3. Fit a curve of the form $Y = bc^X$ to the following data :

Year (X) :	1951	1952	1953	1954	1955	1956	1957
Production in tons (Y) :	201	263	314	395	427	504	612

4. In an experiment in which the growth of duck weed under certain con-

ditions was measured, the following results were obtained :

Weeks (X) ...	0	1	2	3	4	2	6	7	8
No. of friends (Y) ...	20	30	52	77	135	211	326	550	1052

Assuming the relationship of the form $Y = ae^{bX}$, find the best values of a and b by the method of least squares.

5. For the data given below, find the equation to the best fitting exponential curve of the form $Y = ae^{bX}$.

X :	1	2	3	4	5	6
Y :	1·6	4·5	13·8	40·2	125·0	300·0

Ans. $Y = (0{\cdot}557)\, e^{1{\cdot}05X}$

6. Fit the curve $Y = aX^2 + (b/X)$ to the following data :

X :	1	2	3	4
Y :	$-1{\cdot}51$	0·99	3·88	7·66

7. The following table gives corresponding values of two variables X and Y.

X :	1	2	3	4	5
Y :	1·8	5·1	8·9	14·1	19·8

It is found that they are connected by a law of the form $Y = aX + bX^2$, where a and b are constants. Find the best values of a and b by the method of least squares. Calculate the value of Y for $X = 2$.

Ans. $a = 1{\cdot}521$; $b = 0{\cdot}49$; $5{\cdot}006$

8. The following pairs of observations were noted in experimental work on cosmic rays. Find, by the method of least squares, the best values of a and b for the equation $\log R = a - bC$ which fits the data and estimate the most probable value of R for $C = 20{\cdot}7$.

C :	14	15	16	17	18
R :	24·1	20·5	14·0	7·3	5·0

9. *(a)* Explain the principle of least squares and describe its applications in fitting a curve of the form $Y = a \exp{(bX + cX^2)}$.

(b) Fit an indifference curve of the type $XY = b + aX$ to the data given below:

Consumption of Commodity X :	1	2	3	4
Consumption of Commodity Y :	3	1·5	6	7·5

Hint. $y = a + (b/x)$. Now proceed as in Example 9·6.

Ans. $XY = 1{\cdot}3X + 1{\cdot}7$

10. *(a)* Show that the parabola of best fit for the points

$$(x_1, y_1) \,;\, (x_2, y_2) \;;\; \ldots \ldots \;;\, (x_{2n+1}, y_{2n+1})$$

where the values of x are in A.P. with common difference unity and $\bar{x} = 0$, can be expressed in the form

$$\begin{vmatrix} y & x^2 & x & 1 \\ \Sigma y_i & \frac{n(n+1)(2n+1)}{3} & 0 & an+1 \\ \Sigma x_i y_i & 0 & \frac{n(n+1)(2n+1)}{3} & 0 \\ \Sigma x_i^2 y_i & \frac{n^2(n+1)^2}{2} & n & \frac{n(n+1)(2n+1)}{3} \end{vmatrix} = 0$$

[Delhi Univ. B.A. (Pass), 1984]

Hint. Use (9·4*b*), with

$x_i = a + i$; $i = 1, 2, ..., (2n + 1)$. Since $\bar{x} = 0$,

$$\Sigma x_i = (2n + 1)a + \Sigma i \quad \Rightarrow \quad 0 = (2n + 1)a + \frac{(2n + 1)(2n + 2)}{2}$$

$$\Rightarrow \quad a = -(n + 1)$$

$$\Sigma x_i^2 = \Sigma (a + i)^2 = (2n + 1)a^2 + \Sigma i^2 + 2a\, \Sigma i$$

and so on, for Σx_i^3 and Σx_i^4.

$$\left[\sum_{i=1}^{m} i^2 = \frac{m(m + 1)(2m + 1)}{6} ; \quad \sum_{i=1}^{m} i^3 = \left[\frac{m(m + 1)}{2} \right]^2 \right.$$

and $\quad \left. \sum_{i=1}^{m} i^4 = \frac{1}{30} m(m + 1)(2m + 1)(3m^2 + 3m - 1) \right]$

(*b*) When do we prefer logarithmic curve to ordinary curve ?

9·5. Curve Fitting by Orthogonal Polynomials. Suppose that the polynomial of *p*th degree of Y on X is

$$Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_p X^p \qquad ...(9·13)$$

The normal equations for determining the constants a_i's are obtained by the principle of least squares by minimising the residual or error sum of squares

$$E = \Sigma (y - a_0 - a_1 x - a_2 x^2 - \ldots - a_p x^p)^2 \qquad ...(9·14)$$

summation being extended over the given set of observations. The normal equations are :

$$\frac{\partial E}{\partial a_j} = 0, \ (j = 0, 1, 2, ..., p)$$

i.e., $\Sigma x^j (y - a_0 - a_1 x - a_2 x^2 - \ldots - a_p x^p) = 0, \ [j = 0, 1, 2, ..., p]$ ...(9·15)

Assume that X and Y are measured from their means (and this we can do without any loss of generality) so that

$$\mu_r = \mu_r' = E(X^r) = \frac{1}{N} \Sigma x^r$$

and write,

$$\mu_{j1} = \frac{1}{N} \Sigma x^j . y ,$$

where N is number of observations taken on each of the variables X and Y. Hence (9·15) gives

$$\mu_{j1} - a_0 \mu_j - a_1 \mu_{j+1} - a_2 \mu_{j+2} - \ldots - a_p \mu_{j+p} = 0 \; ; \quad j = 0, 1, 2, \ldots, p$$

$$\Rightarrow \quad a_0 \mu_j + a_1 \mu_{j+1} + a_2 \mu_{j+2} + \ldots + a_p \mu_{j+p} = \mu_{j1} \; ; \; j = 0, 1, 2, \ldots, p$$

Putting $j = 0, 1, 2, \ldots, p$, we get respectively

$$\left.\begin{array}{l} a_0 \mu_0 + a_1 \mu_1 + a_2 \mu_2 + \ldots + a_p \mu_p = \mu_{01} \\ a_0 \mu_1 + a_1 \mu_2 + a_2 \mu_3 + \ldots + a_p \mu_{p+1} = \mu_{11} \\ \quad \vdots \qquad \vdots \qquad \vdots \qquad \qquad \vdots \qquad \vdots \\ a_0 \mu_p + a_1 \mu_{p+1} + a_2 \mu_{p+2} + \ldots + a_p \mu_{2p} = \mu_{p1} \end{array}\right\} \quad \ldots(9 \cdot 16)$$

Solving (9·16) for $a_0, a_1, \ldots, a_p$ in terms of the moments μ_j's and μ_{j1}'s, $j = 0, 1, 2, \ldots, p$ and substituting in (9·13) we get the required curve of best fit.

Let

$$\Delta^{(p)} = \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \ldots & \mu_p \\ \mu_1 & \mu_2 & \mu_3 & \ldots & \mu_{p+1} \\ \vdots & \vdots & \vdots & & \vdots \\ \mu_p & \mu_{p+1} & \mu_{p+2} & \ldots & \mu_{2p} \end{vmatrix} \quad \ldots(9 \cdot 17)$$

and $\Delta_j^{(p)}$ be the determinant obtained on replacing $(j+1)$th column of $\Delta^{(p)}$ by the column

$$\begin{pmatrix} \mu_{01} \\ \mu_{11} \\ \vdots \\ \mu_{p1} \end{pmatrix} \quad \text{then} \quad a_j = \frac{\Delta_j^{(p)}}{\Delta^{(p)}} \quad \ldots(9 \cdot 18)$$

The required curve of best fit is the eliminant of a_i's in (9·13) and (9·16) and is given by

$$\begin{vmatrix} Y & 1 & X & X^2 & \ldots & X^p \\ \mu_{01} & \mu_0 & \mu_1 & \mu_2 & \ldots & \mu_p \\ \mu_{11} & \mu_1 & \mu_2 & \mu_3 & \ldots & \mu_{p+1} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \mu_{p1} & \mu_p & \mu_{p+1} & \mu_{p+2} & \ldots & \mu_{2p} \end{vmatrix} = 0 \quad \ldots(9 \cdot 19)$$

The use of equation (9·19) is subject to one serious drawback. If we have a set of data and apart from inspection if there is no guide regarding the order of the polynomial to be fitted, the only way left to us is to try curves of order 1, 2, 3, ... until we reach the point where further terms do not improve the fit. Every time we add a new term, the a_j's given by (9·18) change and accordingly the determinantal arithmetic has to be done afresh. For example, if we want to fit a polynomial curve of third or higher degree to the same data then we cannot use the coefficients which we computed while fitting a second degree parabola. To overcome this drawback Prof. R.A. Fisher suggested a method which involved the fitting of *Orthogonal Polynomials* by the principle of least squares, so that each term is independent of the other, *i.e.*, each of the coefficients in the polynomial is independent of the other so that each of them can be calculated independently. In this method, the coefficients computed earlier remain the same and we have to compute the coefficient only for the added term.

9·5·1. **Orthogonal Polynomials** *(Def.)*. Two polynomials $P_1(x)$ and $P_2(x)$ are said to be *orthogonal* to each other if

$$\Sigma P_1(x)\, P_2(x) = 0, \qquad \text{...(9·20)}$$

where summation is taken over a specified set of values of x. If x were a continuous variable in the range from a to b, the condition for orthogonality gives

$$\int_a^b P_1(x)\, P_2(x)\, dx = 0 \qquad \text{...(9·20a)}$$

For example, if we take

$$P_0 = 1, P_1(x) = x - 4, P_2(x) = x^2 - 8x + 12, P_3(x) = x^3 - 12x^2 + 41x - 36 \qquad \text{...(9·20b)}$$

then these are orthogonal to each other for a set of integral values of x from 1 to 7 as explained in the following table. Other examples of orthogonal polynomials are Hermite polynomials, Gram Charlier's polynomials, Legender's polynomials, etc.

ORTHOGONALITY OF POLYNOMIALS DEFINED IN (9·20b)

x	P_0P_1	P_0P_2	P_0P_3	P_1P_2	P_1P_3	P_2P_3
1	−3	5	−6	−15	18	−30
2	−2	0	6	0	−12	0
3	−1	−3	6	3	−6	−18
4	0	−4	0	0	0	0
5	1	−3	−6	−3	−6	18
6	2	0	−6	0	−12	0
7	3	5	6	15	18	30
Total	0	0	0	0	0	0

9·5·2. **Fitting of Orthogonal Polynomials.** The pth degree polynomial (9·13) can be rewritten as

$$Y = b_0P_0 + b_1P_1 + b_2P_2 + \ldots + b_pP_p \qquad \text{...(9·21)}$$

where P's are polynomials in x, P_j being a polynomial of degree j, $(j = 0, 1, 2, \ldots, p)$. We shall determine P's so that they satisfy the condition of orthogonality, *viz.*,

$$\Sigma P_j \cdot P_k = \sum_x P_j(x)\, P_k(x) = 0\,;\ j \neq k \qquad \text{...(9·22)}$$

the summation being extended over the observed values of x. The normal equations for estimating the constants b_j's are obtained on minimising

$$E = \Sigma\,(y - b_0P_0 - b_1P_1 - \ldots - b_pP_p)^2 \qquad \text{...(9·23)}$$

and are given by

$$\frac{\partial E}{\partial b_j} = 0$$

$$\Rightarrow \quad \Sigma P_j (y - b_0 P_0 - b_1 P_1 - \ldots - b_p P_p) = 0 : j = 0, 1, 2, \ldots, p.$$

Simlifying and using (9·22), we get

$$\Sigma P_j . y - b_j \Sigma P_j^2 = 0$$

$$\Rightarrow \quad b_j = \frac{\Sigma y P_j}{\Sigma P_j^2}, j = 0, 1, 2, \ldots, p. \qquad \ldots(9.24)$$

Thus b_j is determined by P_j. If having fitted a curve of order p we wish to go a step further by adding a term $b_{p+1} P_{p+1}$, the coefficients already obtained in (9·24) remain unaltered.

Moreover, the use of orthogonal polynomials will give us a very convenient method of determining, step by step, the goodness of fit of the polynomial curve. For pth degree polynomial (9·21), the error sum of squares is [c.f. (9·23)]

$$\begin{aligned} E &= \Sigma (y - b_0 P_0 - b_1 P_1 - \ldots - b_p P_p)^2 \\ &= \Sigma y^2 + b_0^2 \Sigma P_0^2 + b_1^2 \Sigma P_1^2 + \ldots + b_p^2 \Sigma P_p^2 \\ &\quad - 2 b_0 \Sigma y P_0 - 2 b_1 \Sigma y P_1 - \ldots - 2 b_p \Sigma y P_p, \end{aligned}$$

other terms vanish because of orthogonality conditions (9·22). Using (9·24) we finally obtain

$$E = \Sigma y^2 - b_0^2 \Sigma P_0^2 - b_1^2 \Sigma P_1^2 - \ldots - b_p^2 \Sigma P_p^2 \qquad \ldots(9.25)$$

Thus the effect of adding any term $b_j P_j$ is to reduce the error (residual) sum of squares E by $b_j^2 \Sigma P_j^2$ and we may examine the effect of this term on E separately. If we find that the addition of any term $b_p P_p$ does not reduce E significantly, we may conclude that it is not desired (as far as the representation of the given data by a polynomial curve is concerned).

9·5·3. Finding The Orthogonal Polynomial P_p, in Fitting a Polynomial of Degree p. Let P_p, the polynomial of degree p in x be given by

$$P_p = \sum_{j=0}^{p} c_{pj} x^j \qquad \ldots(9.26)$$

This contains $(p+1)$ unknown constants $c_{p0}, c_{p1}, \ldots, c_{pp}$. Hence in all the polynomials in (9·21) up to and including those of pth order, there are

$$1 + 2 + 3 + \ldots + (p+1) = \frac{(p+1)(p+2)}{2},$$

unknown constants. The orthogonality conditions

$$\Sigma P_i P_j = 0, i \neq j = 0, 1, 2, \ldots, p,$$

provide ${}^{p+1}C_2 = \frac{(p+1)p}{2}$ conditions on the c's so that there are

$$\frac{(p+1)(p+2)}{2} - \frac{(p+1)p}{2} = p + 1,$$

constants which can be assigned arbitrarily. We will take one for each polynomial P_j (j = 0, 1, 2, ..., p) and assign it such that the coefficient of x^j in P_j is unity i.e.,

$$c_{jj} = 1, j = 0, 1, 2, ..., p, \quad ...(9.27)$$

In particular $c_{00} = P_0 = 1$. The orthogonality conditions give:

$$\sum_x P_p P_j = 0, j < p \ (j = 0, 1, 2, ..., p-1) \quad ...(9·28)$$

$j = 0$, gives $\Sigma P_p P_0 = 0 \Rightarrow \Sigma P_p = 0; (\because P_0 = 1)$ ...(*)

$j = 0$, gives $\Sigma P_p P_1 = 0 \Rightarrow \Sigma P_p = 0, (x + k) = 0$

$\Rightarrow \Sigma P_p \,.\, x + k \Sigma P_p = 0$

$\Rightarrow \Sigma x P_p = 0$ ...(**)

$j = 2$, gives $\Sigma P_p P_2 = 0 \Rightarrow \Sigma P_p = (x^2 + k_1 x + k_2) = 0$ [Using (*)]

$\Rightarrow \Sigma x^2 . P_p = 0$ [Using (*) and (**)]

Similarly proceeding, we shall get in general

$$\sum_x P_p x^r = 0, \quad r = 0, 1, 2, ..., p-1 \quad ...(9·29)$$

$$\Rightarrow \quad \sum_x \left(\sum_{j=0}^{p} c_{pj} \,.\, x^j \right) x^r = 0$$

$$\Rightarrow \quad \sum_{j=0}^{p} \left(c_{pj} \sum_x x^{j+r} \right) = 0$$

Dividing both sides by N, the number of observations on each of the variables X and Y, we get.

$$\sum_{j=0}^{P} c_{pj} \ \mu_{j+r} = 0; r = 0, 1, 2, ..., (p-1) \quad ...(9·30)$$

where x is assumed to be measured from mean. Putting $r = 0, 1, 2, .., (p-1)$ in (9·30), we get respectively

$$c_{p0}\,\mu_0 + c_{p1}\,\mu_1 + \cdots + c_{pj}\,\mu_j + \cdots + c_{p,p-1}\,\mu_{p-1} + c_{pp}\,\mu_p = 0$$
$$c_{p0}\,\mu_1 + c_{p1}\,\mu_2 + \cdots + c_{pj}\,\mu_{j+1} + \cdots + c_{p,p-1}\,\mu_p + c_{pp}\,\mu_{p+1} = 0$$
$$\vdots$$
$$c_{p0}\,\mu_{p-1} + c_{p1}\,\mu_p + \cdots + c_{pj}\,\mu_{j+p-1} + \cdots + c_{p,p-1}\,\mu_{2p-2} + c_{pp}\,\mu_{2p-1} = 0$$

Noting that $c_{pp} = 1$, solving the above equations for c's, we get

$$c_{pj} = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \cdots & -\mu_p & \cdots & \mu_{p-1} \\ \mu_1 & \mu_2 & \cdots & -\mu_{p+1} & \cdots & \mu_p \\ \vdots & \vdots & & \vdots & & \vdots \\ \mu_{p-1} & \mu_p & \cdots & \mu_{2p-1} & \cdots & \mu_{2p-2} \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_j & \cdots & \mu_{p-1} \\ \mu_1 & \mu_2 & \cdots & \mu_{j+1} & \cdots & \mu_p \\ \vdots & \vdots & & \vdots & & \vdots \\ \mu_{p-1} & \mu_p & \cdots & \mu_{j+p-1} & \cdots & \mu_{2p-2} \end{vmatrix}} = \frac{\Delta^{(p)}{}_{pj}}{\Delta^{(p-1)}} \quad ...(9·31)$$

where $\Delta^{(p)}$ has been defined in (9·17) and $\Delta^{(p)}{}_{pj}$ is the minor of the element in the last row and $(j+1)$th column in $\Delta^{(p)}$. Substituting this value of c_{pj} in (9·26), we get

$$P_p = \sum_{j=0}^{p} c_{pj}\, x^j = \sum_{j=0}^{p} \frac{\Delta^{(p)}{}_{pj}}{\Delta^{(p-1)}} \cdot x^j$$

$$= \frac{1}{\Delta^{(p-1)}} \sum_{j=0}^{p} \Delta^{(p)}{}_{pj}\, x^j$$

$$= \frac{1}{\Delta^{(p-1)}} \left[\Delta^{(p)}{}_{p0} + x\,\Delta^{(p)}{}_{p1} + \ldots + x^p \cdot \Delta^{(p)}{}_{pp} \right]$$

$$\Rightarrow \qquad P_p = \frac{1}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_p \\ \mu_1 & \mu_2 & \mu_3 & \mu_{p+1} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{p-1} & \mu_p & \mu_{p+1} & \mu_{2p-1} \\ 1 & x & x^2 & x^p \end{vmatrix} \qquad \ldots(9·32)$$

In particular if $\mu_0 = 1, \mu_1 = 0$ and $\mu_2 = 1$, *i.e.,* if x is a standardised variate then the orthogonal polynomials are given by

$$P_0 = 1 \qquad \ldots(9·33)$$

$$P_1(x) = \frac{\begin{vmatrix} \mu_0 & \mu_1 \\ 1 & x \end{vmatrix}}{\mu_0} = x \qquad \ldots(9·33a)$$

$$P_2(x) = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ 1 & x & x^2 \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{vmatrix}} = x^2 - \mu_3 x - 1 \qquad \ldots(9·33b)$$

$$(\because \mu_0 = 1, \mu_1 = 0, \mu_2 = 1)$$

$$P_3(x) = \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_3 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ \mu_2 & \mu_3 & \mu_4 & \mu_5 \\ 1 & x & x^2 & x^3 \end{vmatrix} \div \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix} \qquad \ldots(9·33c)$$

and so on.

If we further assume that x is a standard *normal* variate so that $\mu_3 = \mu_5 = \ldots = \mu_{2r+1} = 0$, then the above orthogonal polynomials are called *Hermite Polynomials* and are given by

$P_0 = 1$; $P_1(x) = x$; $P_2(x) = x^2 - 1$; $P_3(x) = x^3 - 3x$; $P_4(x) = x^4 - 6x^2 + 3$;
and so on, where x is a continuous r.v. taking values from $-\infty$ to ∞. ...(9·34)

Remark. Hermite Polynomials defined in (9·34) are orthogonal w.r.t. the weight function

$$\alpha(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right); \; -\infty < x < \infty$$

i.e., $$\int_{-\infty}^{\infty} P_i(x)\, P_j(x)\, \alpha(x) = 0 \; ; \; i \neq j \qquad \text{...(9·35)}$$

where $P_1(x), P_2(x), P_3(x), P_4(x)$ are defined in (9·34).

9·5·4. Determination of the Coefficients b_j's in (9·21). From (9·24), we get

$$b_p = \Sigma\, y\, P_p \,/\, \Sigma\, P_p^{\,2} \qquad \text{...(9·36)}$$

Now
$$\Sigma\, P_p^{\,2} = \Sigma\, P_p\, P_p$$
$$= \sum_x P_p\, [c_{p0} + c_{p1}x + c_{p2}x^2 + \ldots + c_{pp}x^p]$$
$$= \sum_x P_p \,.\, x^p,$$

on using (9·29) and the fact that $c_{pp} = 1$.

$$\Sigma\, P_p^{\,2} = \sum_x \left(\sum_{j=0}^{p} c_{pj}\, x^j \right) x^p = \sum_{j=0}^{p} \left(c_{pj} \sum_x x^{p+j} \right)$$

$$= N \sum_{j=0}^{p} c_{pj}\; \mu_{p+j} = N \sum_{j=0}^{p} \frac{\Delta^{(p)}{}_{pj}}{\Delta^{(p-1)}} \cdot \mu_{p+j} \qquad \text{[From (9·31)]}$$

$$= \frac{N}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \ldots & \mu_p \\ \mu_1 & \mu_2 & \ldots & \mu_{p+1} \\ \vdots & & & \\ \mu_{p-1} & \mu_p & \ldots & \mu_{2p-1} \\ \mu_p & \mu_{p+1} & \ldots & \mu_{2p} \end{vmatrix}$$

[Proceeding exactly as we obtained (9·32)]

$$= \frac{N\, \Delta^{(p)}}{\Delta^{(p-1)}} \qquad \text{...(9·37)}$$

Similarly, $\displaystyle \Sigma\, y\, P_p = N \sum_{j=0}^{p} \frac{\Delta^{(p)}{}_{pj}}{\Delta^{(p-1)}} \cdot \mu_{j1}$

$$= \frac{N}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \ldots & \mu_p \\ \mu_1 & \mu_2 & \ldots & \mu_{p+1} \\ \vdots & & & \\ \mu_{p-1} & \mu_p & \ldots & \mu_{2p-1} \\ \mu_{01} & \mu_{11} & \ldots & \mu_{p1} \end{vmatrix}$$

$$= \frac{N \,.\, \Delta^{(p)}}{\Delta^{(p-1)}} \qquad \text{...(9·38)}$$

where $\Delta^{(p)}$ and $\Delta^{(p)}{}_j$ are defined in (9·17). Substituting in (9·36) we get

...(9·39)

If the variable x takes the integral values 1, 2, ..., N, then the first seven of these orthogonal polynomials P_j's $j = 0, 1, 2, 3, ..., 6$ are given by :

$$P_0(x) = 1, P_1(x) = \lambda_1 . \xi$$

$$P_2(x) = \lambda_2 \left\{ \xi^2 - \frac{N^2 - 1}{12} \right\}$$

$$P_3(x) = \lambda_3 \left\{ \xi^3 - \frac{3N^2 - 7}{20} \xi \right\}$$

$$P_4(x) = \lambda_4 \left\{ \xi^4 - \frac{3N^2 - 13}{14} \xi^2 + \frac{3}{560} (N^2 - 1)(N^2 - 9) \right\}$$

$$P_5(x) = \lambda_5 \left\{ \xi^5 - \frac{5}{18} (N^2 - 7) \xi^3 + \frac{1}{1008} (15N^4 - 230N^2 + 407) \xi \right\}$$

$$P_6(x) = \lambda_6 \left\{ \xi^6 - \frac{5}{44} (3N^2 - 31) \xi^4 + \frac{1}{176} (5N^4 - 110N^2 + 329) \xi^2 - \frac{5}{14784} (N^2 - 1)(N^2 - 9)(N^2 - 25) \right\}$$

and so on, where $\xi = x - \bar{x}$ so that $\Sigma \xi = 0$ and λ_i's are arbitrary constants.

If $y = b_0 + b_1 P_1(x) + b_2 P_2(x) + \dots + b_p P_p(x)$; is the orthogonal polynomial fitted to the given data then, using (9·24), we get

$$\left. \begin{aligned} b_0 &= \frac{\Sigma y P_0}{\Sigma P_0^2} = \frac{\Sigma y}{N} \; ; \; (\because P_0 = 1) \\ b_i &= \frac{\sum_x y P_i}{\sum_x P_i^2}, \; (i = 1, 2, ..., p) \end{aligned} \right\} \quad ...(9·40)$$

The origin of P_i's is so chosen that $\Sigma P_i = 0$.

If N, the number of observations is odd, then we take

$$\xi = \frac{x_j - A}{h}$$

and if N is even then we take

$$\xi = \frac{x_i - A_1}{(h/2)}$$

where h = length of the interval (for values of x)

A = middle value (item) of the data

and A_1 = Arithmetic mean of two middle values of the data.

The values of P_i's and λ_i's are obtained from 'Statistical Tables' by

R.A. Fisher for the values of N from 3 to 75. In these tables the orthogonal polynomials P_i's are denoted by ϕ_i's. We reproduce below these tables for $N = 3$ to $N = 6$.

TABLES OF ORTHOGONAL POLYNOMIALS

	N = 3		N = 4			N = 5			
	φ_1	φ_2	φ_1	φ_2	φ_3	φ_1	φ_2	φ_3	φ_4
	–1	1	–3	1	–1	–2	2	–1	1
	0	–2	–1	–1	–3	–1	–1	2	–4
	1	1	1	–1	–3	0	–2	0	6
			3	1	1	1	–1	–2	–4
						2	2	1	1
$\sum_x \phi_i^2$	2	6	20	4	20	10	14	10	70
λ_i	1	3	2	1	$\frac{10}{3}$	1	1	$\frac{5}{6}$	$\frac{35}{22}$

	N = 6				
	φ_1	φ_2	φ_3	φ_4	φ_5
	–5	5	–5	1	–1
	–3	–1	7	–3	5
	–1	–4	4	2	–10
	1	–4	–4	–2	10
	3	–1	–7	–3	–5
	5	5	5	1	1
$\sum_x \varphi_i^2$	70	84	180	28	252
λ_i	2	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{7}{12}$	$\frac{21}{10}$

Example 9·8. *Fit a straight line $y = a + bx$...(*) to the following data by using orthogonal polynomials.*

x	0	1	2	3	4
y	1	1·8	3·3	4·5	6·3

Solution. Here $N = 5$. Let us transform to the variable

$$\xi = \frac{x-2}{1} = x - 2 \text{ so that } \Sigma\xi = 0$$

Let the orthogonal polynomial form of straight line (*) be

$$y = b_0 + b_1 P_1(x) = b_0 + b_1 \phi_1(x) \qquad \text{...(**)}$$

x	$\xi = x - 2$	y	ϕ_1	$y\,\phi_1$
0	−2	1	−2	−2
1	−1	1·8	−1	−1·8
2	0	3·3	0	0
3	1	4·5	1	4·5
4	2	6·3	2	12·6
Total		16·9	·0	13·3

The values of ϕ_1 are noted from the tables for $N = 5$. From tables we also find

$$\Sigma\,\phi_1^2 = 10,\ \lambda_1 = 1$$

Now using (9·40), $b_0 = \dfrac{\Sigma y}{N} = \dfrac{16\cdot 9}{5} = 3\cdot 38$; $b_1 = \dfrac{\Sigma\, y\,\phi_1}{\Sigma\,\phi_1^2} = \dfrac{13\cdot 3}{10} = 1\cdot 33$

$$\phi_1(x) = \lambda_1\,\xi = 1\,.\,(x-2) = x - 2$$

Substituting in (**), the required straight line is

$$y = 3\cdot 38 + 1\cdot 33\,(x-2)$$

$$\Rightarrow \qquad y = 1\cdot 33x + 0\cdot 72$$

Example 9·9. *Fit a second degree parabola to the following data, using the method of orthogonal polynomials.*

x	0·5	1·0	1·5	2·0	2·5	3·0
y	72	110	158	214	290	380

Solution. Let the second degree parabola be

$$y = a + bx + cx^2 \qquad \ldots(*)$$

and its orthogonal polynomial transform be :

$$y = b_0 + b_1\,\phi_1(x) + b_2\,\phi_2(x) \qquad \ldots(**)$$

Here we have $N = 6$. Let us transform to

$$\xi = \frac{x - \frac{1}{2}(1\cdot 5 + 2\cdot 0)}{\frac{1}{2}(0\cdot 5)} = 4\,(x - 1\cdot 75) = 4x - 7,$$

so that $\Sigma\,\xi = 0$. From Fisher's tables we note the values of ϕ_1 and ϕ_2 (as given in the following table) and also

$\Sigma\,\phi_1^2 = 70,\ \Sigma\,\phi_2^2 = 84\ ;\ \lambda_1 = 2,\ \lambda_2 = 3/2$

x	$\xi = 4x - 7$	y	ϕ_1	ϕ_2	$y\,\phi_1$	$y\,\phi_2$
0·5	−5	72	−5	5	−360	360
1·0	−3	110	−3	−1	−330	−110

1·5	–1	158	–1	–4	–158	–632
2·0	1	214	1	–4	214	–856
2·5	3	290	3	–1	870	–290
3·0	5	380	5	5	1900	1900
Total		1224			2136	372

$$b_0 = \frac{\Sigma y}{N} = \frac{1224}{6} = 204\ ;\ b_1 = \frac{\Sigma y\,\phi_1}{\Sigma \phi_1^2} = \frac{2136}{70} = 30{\cdot}51$$

$$b_2 = \frac{\Sigma y\,\phi_2}{\Sigma \phi_2^2} = \frac{372}{84} = 4{\cdot}43\ ;\ \phi_1(x) = \lambda_1 \xi = 2[4x - 7] = 8x - 14$$

$$\phi_2(x) = \lambda_2\left[\xi^2 - \frac{N^2 - 1}{12}\right] = \frac{3}{2}\left[(4x-7)^2 - \frac{36-1}{12}\right]$$

$$= \frac{3}{2}\left[16x^2 + 49 - 56x - \frac{35}{12}\right] = 24x^2 - 84x + 69{\cdot}125$$

Substituting in (**), we get

$$y = 204 + 30{\cdot}51(8x - 14) + 4{\cdot}43\,(24x^2 - 84x + 69{\cdot}125)$$
$$= 106{\cdot}32x^2 - 128{\cdot}04x + 83{\cdot}08$$

which is the required second degree parabola of best fit.

Correlation and Regression

10·1. Bivariate Distribution, Correlation. So far we have confined ourselves to unvariate distributions, *i.e.*, the distributions involving only one variable. We may, however, come across certain series where each term of the series may assume the values of two or more variables. For example, if we measure the heights and weights of a certain group of persons, we shall get what is known as *Bivariate distribution*—one variable relating to height and other variable relating to weight.

In a bivariate distribution we may be interested to find out if there is any correlation or covariation between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, *i.e.*, if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct* or *positive*. But if they constantly deviate in the opposite directions, *i.e.*, if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be *diverse* or *negative*. For example, the correlation between (*i*) the heights and weights of a group of persons, (*ii*) the income and expenditure is positive and the correlation between (*i*) price and demand of a commodity, (*ii*) the volume and pressure of a perfect gas, is negative. Correlation is said to be *perfect* if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

10·2. Scatter Diagram. It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$, if the values of the variables X and Y be plotted along the x-axis and y-axis respectively in the xy plane, the diagram of dots so obtained is known as *scatter diagram*. From the scatter diagram, we can form a fairly good, though vague, idea whether the variables are correlated or not, *e.g.*, if the points are very dense, *i.e.*, very close to each other, we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. This method, however, is not suitable if the number of observations is fairly large.

10·3. Karl Pearson Coefficient of Correlation. As a measure of itensity or degree of linear relationship between two variables, Karl Pearson (1867—1936), *a British Biometrician*, developed a formula called *Correlation Coefficient*.

Correlation coefficient between two random variables X and Y, usually denoted by $r(X, Y)$ or simply r_{XY}, is a numerical measure of *linear relationship* between them and is defined as

$$r(X, Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X \sigma_Y} \qquad (10·1)$$

If (x_i, y_i) ; $i = 1, 2, ..., n$ is the bivariate distribution, then

$$\left.\begin{aligned} \text{Cov}(X, Y) &= E[\{X - E(X)\}\{(Y - E(Y)\}] \\ &= \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \mu_{11} \\ \sigma_X^2 &= E\{X - E(X)\}^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2 \\ \sigma_Y^2 &= E\{Y - E(Y)\}^2 = \frac{1}{n}\Sigma(y_i - \bar{y})^2 \end{aligned}\right\}, \qquad \ldots (10\cdot2)$$

the summation extending over i from 1 to n.

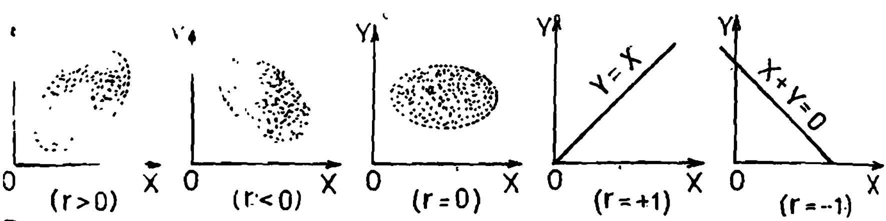
Another convenient form of the formula (10·2) for computational work is as follows :

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\,\bar{y}) \\ &= \frac{1}{n}\Sigma x_i y_i - \bar{y}\frac{1}{n}\Sigma x_i - \bar{x}\frac{1}{n}\Sigma y_i + \bar{x}\,\bar{y} \end{aligned}$$

$$\therefore \qquad \text{Cov}(X, Y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\,\bar{y}, \quad \sigma_X^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2$$

and
$$\sigma_Y^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 \qquad \ldots (10.2a)$$

Remarks 1. Following are the figures of the standard data for $r > 0$, < 0, $= 0$, and $r = \pm 1$.



2. It may be noted that $r(X, Y)$ provides a measure of *linear relationship* between X and Y. For nonlinear relationship, however, it is not very suitable.

3. Sometimes, we write : $\text{Cov}(X, Y) = \sigma_{XY}$

4. Karl Pearson's correlation coefficient is also called *product-moment correlation coefficient,* since

$\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = \mu_{11}$.

10·3·1. Limits for Correlation Coefficient. We have

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma(x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma(y_i - \bar{y})^2\right]^{1/2}},$$

$$\therefore \quad r^2(X, Y) = \frac{(\sum a_i b_i)^2}{(\sum a_i^2)(\sum b_i^2)}, \text{ where } \begin{pmatrix} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{pmatrix} \quad \ldots(*)$$

We have the Schwartz inequality which states that if $a_i, b_i; i = 1, 2, \ldots, n$ are real quantities then

$$(\sum_{i=1}^{n} a_i b_i)^2 \le (\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)$$

the sign of equality holding if and only if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$$

Using Schwartz inequality, we get from (*)

$$r^2(X, Y) \le 1 \text{ i.e., } |r(X, Y)| \le 1 \Rightarrow -1 \le r(X, Y) \le 1 \quad \ldots(10{\cdot}3)$$

Hence correlation coefficient cannot exceed unity numerically. It always lies between -1 and $+1$. If $r = +1$, the correlation is perfect and positive and if $r = -1$, correlation is perfect and negative.

Aliter. If we write $E(X) = \mu_X$ and $E(Y) = \mu_Y$, then we have

$$E\left[\left(\frac{X - \mu_X}{\sigma_X}\right) \pm \left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right]^2 \ge 0$$

$$\Rightarrow \quad E\left(\frac{X - \mu_X}{\sigma_X}\right)^2 + E\left(\frac{Y - \mu_Y}{\sigma_Y}\right)^2 \pm 2\frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \ge 0$$

$$\Rightarrow \quad 1 + 1 \pm 2\, r(X, Y) \ge 0$$

$$\Rightarrow \quad -1 \le r(X, Y) \le 1.$$

Theorem 10·1. *Correlation coefficient is independent of change of origin and scale.*

Proof. Let $U = \dfrac{X - a}{h}$, $V = \dfrac{Y - b}{k}$, so that $X = a + hU$ and $Y = b + kV$, where a, b, h, k are constants; $h > 0, k > 0$.

We shall prove that $r(X, Y) = r(U, V)$

Since $X = a + hU$ and $Y = b + kV$, on taking expectations, we get

$$E(X) = a + hE(U) \quad \text{and} \quad E(Y) = b + kE(V)$$

$$\therefore \quad X - E(X) = h[U - E(U)] \quad \text{and} \quad Y - E(Y) = k[V - E(V)]$$

$$\begin{aligned} \Rightarrow \text{Cov}(X, Y) &= E[\{X - E(X)\}\{Y - E(Y)\}] \\ &= E[h\{U - E(U)\}\, \{k\{V - E(V)\}] \\ &= hk\, E[\{U - E(U)\}\{V - E(V)\}] = hk \text{ Cov}(U, V) \quad \ldots(10{\cdot}4) \end{aligned}$$

$$\sigma_X^2 = E[\{X - E(X)\}^2] = E[h^2\{U - E(U)\}^2] = h^2\sigma_U^2$$

$$\Rightarrow \quad \sigma_X = h\sigma_U, \ (h > 0) \quad \ldots(10{\cdot}4a)$$

and

$$\sigma_Y^2 = E[\{Y - E(Y)\}^2] = E[k^2\{V - E(V)\}^2] = k^2\sigma_V^2$$

$$\Rightarrow \quad \sigma_Y = k\sigma_V, \ (k > 0) \quad \ldots(10{\cdot}4b)$$

Substituting from (10·4), (10·4*a*) and (10·4*b*) in (10·1), we get

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{hk.\ \text{Cov}(U, V)}{hk.\sigma_U \sigma_V} = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = r(U, V)$$

This theorem is of fundamental importance in the numerical computation of the correlation coefficient.

Corollary. *If X and Y are random variables and a, b, c, d are any numbers provided only that $a \neq 0$, $c \neq 0$, then*

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y)$$

Proof. With usual notations, we have

$$\text{Var}(aX + b) = a^2\sigma_X^2\ ;\ \text{Var}(cY + d) = c^2\sigma_Y^2\ ;$$

$$\text{Cov}(aX + b, cY + d) = ac\sigma_{XY}$$

$$\therefore\quad r(aX + b,\ cY + d) = \frac{\text{Cov}(aX + b, cY + d)}{[\text{Var}(aX + b)\ \text{Var}(cY + d)]^{1/2}}$$

$$= \frac{ac\ \sigma_{XY}}{|a|\,|c|\,\sigma_X\ \sigma_Y} = \frac{ac}{|ac|} r(X, Y)$$

If $ac > 0$, *i.e.*, if a and c are of same signs, then $ac/|ac| = +1$

If $ac < 0$, *i.e.*, if a and c are of opposite signs, then $ac/|ac| = -1$.

Theorem 10·2. *Two independent variables are uncorrelated.*

Proof. If X and Y are independent variables, then

$$\text{Cov}(X, Y) = 0 \qquad (cf.\ \S\ 6\cdot4)$$

$$\therefore\quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = 0$$

Hence two independent variables are uncorrelated.

But the converse of the theorem is not true, *i.e.*, two uncorrelated variables may not be independent as the following example illustrates :

X	−3	−2	−1	1	2	3	Total $\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	−27	− 8	−1	1	8	27	$\Sigma XY = 0$

$$\bar{X} = \frac{1}{n}\Sigma X = 0,\ \text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y} = 0$$

$$\therefore\quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\ \sigma_Y} = 0$$

Thus in the above example, the variables X and Y are uncorrelated. But on careful examination we find that X and Y are not independent but they are connected by the relation $Y = X^2$. Hence two uncorrelated variables need not necessarily be independent. A simple reasoning for this strange conclusion is that $r(X, Y) = 0$, merely implies the absence of any linear relationship between the variables X and Y. There may, however, exist some other form of relationship between them, *e.g.*, quadratic, cubic or trigonometric.

Remarks. 1. Following are some more examples where two variables are uncorrelated but *not* independent.

(*i*) $X \sim N(0, 1)$ and $Y = X^2$

Since $X \sim N(0, 1)$, $E(X) = 0 = E(X^3)$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = E(X^3) - E(X)\,E(Y) = 0 \qquad (\because Y = X^2)$$

$$\Rightarrow \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\,\sigma_Y} = 0$$

Hence X and Y are uncorrelated but not independent.

(*ii*) Let X be a r.v. with p.d.f.

$$f(x) = \frac{1}{2}, \; -1 \le x \le 1$$

and let $Y = X^2$. Here we shall get

$$E(X) = 0 \text{ and } E(XY) = E(X^3) = 0, \Rightarrow r(X, Y) = 0$$

2. However, the converse of the theorem holds in the following cases :

(*a*) If X and Y are jointly normally distributed with $\rho = \rho(X, Y) = 0$, then they are independent. If $\rho = 0$, then [*c.f.* § 10·10, Equation (10·25)]

$$f(x, y) = \frac{1}{\sigma_X\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{X - \mu_X}{\sigma_X}\right)^2\right] \times \frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{Y - \mu_Y}{\sigma_Y}\right)^2\right]$$

$$\therefore \quad f(x, y) = f_1(x)\,f_2(y)$$

$\Rightarrow$ X and Y are independent.

(*b*) *If each of the two variables X and Y takes two values, 0, 1 with positive probabilities, then $r(X, Y) = 0 \Rightarrow X$ and Y are independent.*

Proof. Let X take the values 1 and 0 with positive probabilities p_1 and q_1 respectively and let Y take the values 1 and 0 with positive probabilities p_2 and q_2 respectively. Then

$$r(X, Y) = 0 \quad \Rightarrow \quad \text{Cov}(X, Y) = 0$$

$$\begin{aligned}
\Rightarrow \quad 0 &= E(XY) - E(X)E(Y)\\
&= 1 \,.\, P(X = 1 \cap Y = 1) - [1 \,.\, P(X = 1) \times 1 \,.\, P(Y = 1)]\\
&= P(X = 1 \cap Y = 1) - p_1 p_2
\end{aligned}$$

$$\Rightarrow \quad P(X = 1 \cap Y = 1) = p_1 p_2 = P(X = 1)\,.\,P(Y = 1)$$

$\Rightarrow$ X and Y are independent.

10·3·2. Assumptions Underlying Karl Pearson's Correlation Coefficient. Pearsonian correlation coefficient r is based on the following assumptions :

(*i*) *The variables X and Y under study are linearly related.* In other words, the scatter diagram of the data will give a straight line curve.

(ii) *Each of the variables (series) is being affected by a large number of independent contributory causes of such a nature as to produce normal distribution.* For example, the variables (series) relating to ages, heights, weight, supply, price, etc., conform to this assumption. In the words of Karl Pearson :

"The sizes of the complex of organs (something measurable) are determined by a great variety of independent contributory causes, for example, climate, nourishment, physical training and innumerable other causes which cannot be individually observed or their effects measured." Karl Pearson further observes, "The variations in intensity of the contributory causes are small as compared with their absolute intensity and these variations follow the normal law of distribution."

(iii) *The forces so operating on each of the variable series are not independent of each other but are related in a causal fashion.* In other word, cause and effect relationship exists between different forces operating on the items of the two variable series. These forces must be common to both the series. If the operating forces are entirely independent of each other and not related in any fashion, then there cannot be any correlation between the variables under study.

For example, the correlation coefficient between,

(a) the series of heights and incomes of individuals over a period of time,

(b) the series of marriage rate and the rate of agricultural production in a country over a period of time,

(c) the series relating to the size of the shoe and intelligence of a group of individuals,

should be zero, since the forces affecting the two variable series in each of the above cases are entirely independent of each other.

However, if in any of the above cases the value of r for a given set of data is not zero, then such correlation is termed as *chance correlation* or *spurious* or *non-sense correlation.*

Example 10·1. *Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y) :*

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

Solution.

CALCULATIONS FOR CORRELATION COEFFICIENT

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
Total 544	552	37028	38132	37560

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{544}{8} = 68,\ \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{8} \times 552 = 69$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \bar{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left[\frac{37028}{8} - (68)^2\right]\left[\frac{38132}{8} - (69)^2\right]}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628 \cdot 5 - 4624)\ (4766 \cdot 5 - 4761)}} = \frac{3}{\sqrt{4 \cdot 5 \times 5 \cdot 5}} = 0 \cdot 603$$

Aliter.

(SHORT-CUT METHOD)

X	Y	$U = X - 68$	$V = Y - 69$	U^2	V^2	UV
65	67	−3	−2	9	4	6
66	68	−2	−1	4	1	2
67	65	−1	−4	1	16	4
67	68	−1	−1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

$$\bar{U} = \frac{1}{n}\Sigma U = 0,\ \bar{V} = \frac{1}{n}\Sigma V = 0$$

$$\text{Cov}(U, V) = \frac{1}{n}\Sigma UV - \bar{U}\,\bar{V} = \frac{1}{8} \times 24 = 3$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - (\bar{U})^2 = \frac{1}{8} \times 36 = 4 \cdot 5$$

$$\sigma_V^2 = \frac{1}{n}\Sigma V^2 - (\bar{V})^2 = \frac{1}{8} \times 44 = 5 \cdot 5$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{3}{\sqrt{4 \cdot 5 \times 5 \cdot 5}} = 0 \cdot 603 = r(X, Y)$$

Remark. The reader is advised to calculate the correlation coefficient by arbitrary origin method rather than by the direct method, since the latter leads to much simpler arithmetical calculations.

Example 10·2. *A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results :*

$$n = 25,\ \Sigma X = 125,\ \Sigma X^2 = 650,\ \Sigma Y = 100,\ \Sigma Y^2 = 460,\ \Sigma XY = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Obtain the correct value of correlation coefficient.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1988, 1991*]

Solution.

Corrected $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\Sigma XY = 508 - 6\times14 - 8\times6 + 8\times12 + 6\times8 = 520$

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{1}{25}\times 125 = 5,\quad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25}\times 100 = 4$$

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = \frac{1}{25}\times 520 - 5\times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{25}\times 650 - (5)^2 = 1$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{25}\times 436 - 16 = \frac{36}{25}$$

$$\therefore \quad \text{Corrected } r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\,\sigma_Y} = \frac{\frac{4}{5}}{1\times\frac{6}{5}} = \frac{2}{3} = 0{\cdot}67$$

Example 10·3. *Show that if X', Y' are the deviations of the random variables X and Y from their respective means then*

(i) $$r = 1 - \frac{1}{2N}\sum_i\left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2$$

(ii) $$r = -1 + \frac{1}{2N}\sum_i\left(\frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y}\right)^2$$

Deduce that $-1 \le r \le +1$.

[*Delhi Univ. B.Sc. Oct. 1992; Madras Univ. B.Sc., Nov. 1991*]

Solution. (i) Here $X_i' = (x_i - \bar{X})$ and $Y_i' = (Y_i - \bar{Y})$

$$\text{R.H.S.} = 1 - \frac{1}{2N}\sum_i\left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2$$

$$= 1 - \frac{1}{2N} \sum_i \left[\frac{X_i'^2}{\sigma_X^2} + \frac{Y_i'^2}{\sigma_Y^2} - \frac{2X_i'Y_i'}{\sigma_X\sigma_Y} \right]$$

$$= 1 - \frac{1}{2N} \left[\frac{1}{\sigma_X^2} \sum_i X_i'^2 + \frac{1}{\sigma_Y^2} \sum_i Y_i'^2 - \frac{2}{\sigma_X\sigma_Y} \sum_i X_i'Y_i' \right]$$

$$= 1 - \frac{1}{2N} \left[\frac{1}{\sigma_X^2} \sum_i (X_i - \bar{X})^2 + \frac{1}{\sigma_Y^2} \sum_i (Y_i - \bar{Y})^2 - \frac{2}{\sigma_X\sigma_Y} \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) \right]$$

$$= 1 - \frac{1}{2} \left[\frac{1}{\sigma_X^2} \cdot \sigma_X^2 + \frac{1}{\sigma_Y^2} \cdot \sigma_Y^2 - \frac{2}{\sigma_X\sigma_Y} \cdot r\sigma_X\sigma_Y \right]$$

$$= 1 - \frac{1}{2}[1 + 1 - 2r] = r$$

(*ii*) Proceeding similarly, we will get

$$\text{R.H.S.} = -1 + \frac{1}{2}(1 + 1 + 2r) = r$$

Deduction. Since $\left(\frac{X_i'}{\sigma_X} \pm \frac{Y_i'}{\sigma_Y}\right)^2$, being the square of a real quantity is always non-negative, $\sum_i \left(\frac{X_i'}{\sigma_X} \mp \frac{Y_i'}{\sigma_Y}\right)^2$ is also non-negative. From part (*i*), we get

$$r = 1 - \text{(some non-negative quantity)} \Rightarrow r \le 1 \qquad \ldots(*)$$

Also from part (*ii*), we get

$$r = -1 + \text{(some non-negative quantity)} \Rightarrow -1 \le r \qquad \ldots(**)$$

The sign of equality in (*) and (**) holds if and only if

$$\left.\begin{aligned} \frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y} &= 0 \\ \text{and} \quad \frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y} &= 0 \end{aligned}\right\} \forall\, i = 1, 2, \ldots, n$$

respectively.

From (*) and (**), we get

$$-1 \le r \le 1$$

Example 10·4. *The variables X and Y are connected by the equation $aX + bY + c = 0$. Show that the correlation between them is -1 if the signs of a and b are alike and $+1$ if they are different.*

[***Nagpur Univ. B.Sc. 1992; Delhi Univ. B.Sc. (Stat. Hons.) 1992***]

Solution. $aX + bY + c = 0 \Rightarrow aE(X) + bE(Y) + c = 0$

$$\therefore \quad a\{X - E(X)\} + b\{Y - E(Y)\} = 0$$

$$\Rightarrow \quad \{X - E(X)\} = -\frac{b}{a}\{Y - E(Y)\}$$

$$\therefore \quad \text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

$$= -\frac{b}{a} E[\{Y - E(Y)\}^2] = -\frac{b}{a} \cdot \sigma_Y^2$$

$$E\{X - E(X)\}^2 = \frac{b^2}{a^2} E[\{Y - E(Y)\}^2] = \frac{b^2}{a^2} \cdot \sigma_Y^2$$

$$\therefore \quad r = \frac{-\frac{b}{a} \cdot \sigma_Y^2}{\sqrt{\sigma_Y^2}\sqrt{\frac{b^2}{a^2} \cdot \sigma_Y^2}} = \frac{-\frac{b}{a}\sigma_Y^2}{\left|\frac{b}{a}\right| \sigma_Y^2}$$

$$= \begin{cases} +1, & \text{if } b \text{ and } a \text{ are of opposite signs.} \\ -1, & \text{if } b \text{ and } a \text{ are of same sign.} \end{cases}$$

Example 10·5. (a) *If $Z = aX + bY$ and r is the correlation coefficient between X and Y, show that*

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y$$

(b) *Show that the correlation coefficient r between two random variables X and Y is given by*

$$r = (\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2) / 2\sigma_X\sigma_Y$$

where σ_X, σ_Y and σ_{X-Y} are the standard deviations of X, Y and $X - Y$ respectively.

[***Calcutta Univ. B.Sc., 1992; M.S. Baroda Univ. B.Sc. 1992***]

Solution. Taking expectation of both sides of $Z = aX + bY$, we get

$$E(Z) = aE(X) + bE(Y)$$

$$\therefore \quad Z - E(Z) = a\{X - E(X)\} + b\{Y - E(Y)\}$$

Squaring and taking expectation of both sides, we get

$$\begin{aligned} \sigma_Z^2 &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab \text{ Cov } (X, Y) \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y \end{aligned}$$

(*b*) Taking $a = 1$, $b = -1$ in the above case, we have

$$Z = X - Y \text{ and } \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y$$

$$\therefore \quad r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$$

Remark. In the above example, we have obtained

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab \text{ Cov } (X, Y)$$

Similarly, we could obtain the result

$$V(aX - bY) = a^2 V(X) + b^2 V(Y) - 2ab \text{ Cov } (X, Y).$$

The above results are useful in solving theoretical problems.

Example 10·6. *X and Y are two random variables with variances σ_X^2 and σ_Y^2 respectively and r is the coefficient of correlation between them. If $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y} Y$, find the value of k so that U and V are uncorrelated.* [***Delhi Univ. B.Sc. 1992; Andhra Univ. B.Sc. 1993***]

Solution. Taking expectations of $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y} Y$, we get

$$E(U) = E(X) + kE(Y) \text{ and } E(V) = E(X) + \frac{\sigma_X}{\sigma_Y} E(Y)$$

$$U - E(U) = \{X - E(X)\} + k\{Y - E(Y)\} \text{ and}$$

$$V - E(V) = \{X - E(X)\} + \frac{\sigma_X}{\sigma_Y} \{Y - E(Y)\}$$

$$\text{Cov}(U, V) = E[\{U - E(U)\}\{V - E(V)\}]$$

$$= E\Big[\{X - E(X)\} + k(Y - E(Y)\}] \times [\{X - E(X)\} + \frac{\sigma_X}{\sigma_Y}\{Y - E(Y)\}\Big]$$

$$= \sigma_X^2 + \frac{\sigma_X}{\sigma_Y} \text{Cov}(X, Y) + k \text{ Cov}(X, Y) + k\frac{\sigma_X}{\sigma_Y} \cdot \sigma_Y^2$$

$$= \left[\sigma_X^2 + k\sigma_X\sigma_Y\right] + \left[\frac{\sigma_X}{\sigma_Y} + k\right] \text{Cov}(X, Y)$$

$$= \sigma_X(\sigma_X + k\sigma_Y) + \left[\frac{\sigma_X + k\sigma_Y}{\sigma_Y}\right] \text{Cov}(X, Y)$$

$$= (\sigma_X + k\sigma_Y)\left[\sigma_X + \frac{\text{Cov}(X, Y)}{\sigma_Y}\right] = (\sigma_X + k\sigma_Y)(1 + r)\sigma_X$$

U and V will be uncorrelated if

$$r(U, V) = 0 \quad \Rightarrow \quad \text{Cov}(U, V) = 0$$

i.e., if $\quad (\sigma_X + k\sigma_Y)(1 + r)\sigma_X = 0$

$\Rightarrow \quad \sigma_X + k\sigma_Y = 0 \qquad (\because \sigma_X \neq 0, r \neq -1)$

$\Rightarrow \quad k = -\dfrac{\sigma_X}{\sigma_Y}$

Example 10·7. *The random variables X and Y are jointly normally distributed and U and V are defined by*

$$U = X \cos\alpha + Y \sin\alpha,$$
$$V = Y \cos\alpha - X \sin\alpha$$

Show that U and V will be uncorrelated if

$$\tan 2\alpha = \frac{2r\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2},$$

where $r = \text{corr.}(X, Y)$, $\sigma_X^2 = \text{Var}(X)$ *and* $\sigma_Y^2 = \text{Var}(Y)$. *Are U and V then independent?*

[***Delhi Univ. B.Sc. (Stat. Hons.) 1989; (Maths. Hons.), 1990***]

Solution. We have

$$\text{Cov}(U, V) = E[\{U - E(U)\}\{V - E(V)\}]$$

$$= E\Big[[\{X - E(X)\}\cos\alpha + \{Y - E(Y)\}\sin\alpha]$$
$$\times [\{Y - E(Y)\}\cos\alpha - \{X - E(X)\}\sin\alpha]\Big]$$

$$= \cos^2 \alpha \text{ Cov }(X, Y) - \sin \alpha \cos \alpha . \sigma_X^2 + \sin \alpha \cos \alpha . \sigma_Y^2 - \sin^2 \alpha \text{ (Cov }(X, Y)$$
$$= (\cos^2 \alpha - \sin^2 \alpha) \text{ Cov }(X, Y) - \sin \alpha \cos \alpha \; (\sigma_X^2 - \sigma_Y^2)$$
$$= \cos 2\alpha \text{ Cov }(X, Y) - \sin \alpha \cos \alpha \; (\sigma_X^2 - \sigma_Y^2)$$

U and V will be uncorrelated if and only if

$$r(U, V) = 0, \quad i.e., \quad \text{iff Cov }(U, V) = 0$$

i.e., if $\cos 2\alpha \text{ Cov }(X, Y) - \sin \alpha \cos \alpha \; (\sigma_X^2 - \sigma_Y^2) = 0$

or if $\qquad \cos 2\alpha \; r \; \sigma_X \sigma_Y = \dfrac{\sin 2\alpha}{2} . (\sigma_X^2 - \sigma_Y^2)$

or if $\qquad \tan 2\alpha = \dfrac{2r \; \sigma_X \sigma_Y}{\sigma_X^2 - \sigma_Y^2}$

However, $r(U, V) = 0$ does not imply that the variables U and V are independent. [For detailed discussion, see Theorem 10·2, page 10·4.].

Example 10·8. *If X, Y are standardized random variables, and*

$$r(aX + bY, bX + aY) = \frac{1 + 2ab}{a^2 + b^2} \qquad \ldots(*)$$

find r(X, Y), the coefficient of correlation between X and Y.
[***Sardar Patel Univ. B.Sc., 1993; Delhi Univ. B.Sc. (Stat. Hons.), 1989***]

Solution. Since X and Y are standardised random variables, we have

$$\left.\begin{aligned} & E(X) = E(Y) = 0 \\ \text{and} \quad & \text{Var }(X) = \text{Var }(Y) = 1 \Rightarrow E(X^2) = E(Y^2) = 1 \\ \text{and} \quad & \text{Cov}(X, Y) = E(XY) \Rightarrow E(XY) = r(X,Y).\sigma_X\sigma_Y \\ & \qquad = r(X,Y) \end{aligned}\right] \quad \ldots(**)$$

Also we have

$r(aX + bY, bX + aY)$

$$= \frac{E[(aX + bY)(bX + aY)] - E(aX + bY)\, E(bX + aY)}{[\text{Var }(aX + bY) \,.\, \text{Var }(bX + aY)]^{1/2}}$$

$$= \frac{E[abX^2 + a^2 XY + b^2 YX + abY^2] - 0}{\{[a^2 \text{ Var }(X) + b^2 \text{ Var }(Y) + 2ab \text{ Cov }(X,Y)] \times [b^2 \text{ Var }(X) + a^2 \text{ Var } Y + 2ba \text{ Cov }(X,Y)]\}^{1/2}}$$

$$= \frac{ab.1 + a^2 r(X, Y) + b^2 r(X, Y) + ab.1}{\{[a^2 + b^2 + 2ab\, r(X, Y)][b^2 + a^2 + 2ba\, r(X, Y)]\}^{1/2}}$$

[Using (**)]

$$= \frac{2ab + (a^2 + b^2).\, r(X, Y)}{a^2 + b^2 + 2ab.\, r(X, Y)}$$

From (*) and (**), we get

$$\frac{1 + 2ab}{a^2 + b^2} = \frac{(a^2 + b^2).\, r(X, Y) + 2ab}{a^2 + b^2 + 2ab.\, r(X, Y)}$$

Cross multiplying, we get

$$(a^2+b^2)(1+2ab)+2ab.\,r(X,Y)(1+2ab)=(a^2+b^2)^2.\,r(X,Y)+2ab\,(a^2+b^2)$$

$$\Rightarrow \quad (a^4+b^4+2a^2b^2-2ab-4a^2b^2).\,r(X,Y)=(a^2+b^2)$$

$$\Rightarrow \quad [(a^2-b^2)^2-2ab]\,r(X,Y)=a^2+b^2$$

$$\Rightarrow \quad r(X,Y)=\frac{a^2+b^2}{(a^2-b^2)^2-2ab}$$

Example 10·9. *If X and Y are uncorrelated random variables with means zero and variances σ_1^2 and σ_2^2 respectively, show that*

$$U=X\cos\alpha+Y\sin\alpha,\ V=X\sin\alpha-Y\cos\alpha$$

have a correlation coefficient ρ *given by*

$$\rho=\frac{\sigma_1^2-\sigma_2^2}{[(\sigma_1^2-\sigma_2^2)^2+4\sigma_1^2\sigma_2^2\operatorname{cosec}^2 2\alpha]^{1/2}}$$

Solution. We are given that

$$r(X,Y)=0\Rightarrow \text{Cov}(X,Y)=0,\ \sigma_X^2=\sigma_1^2 \text{ and } \sigma_Y^2=\sigma_2^2 \qquad \ldots(1)$$

We have

$$\begin{aligned}\sigma_U^2 &= V(X\cos\alpha+Y\sin\alpha)\\ &=\cos^2\alpha\,V(X)+\sin^2\alpha\,V(Y)+2\sin\alpha\cos\alpha\,\text{Cov}(X,Y)\\ &=\cos^2\alpha\,\sigma_1^2+\sin^2\alpha\,\sigma_2^2 \qquad \text{[Using (1)]}\end{aligned}$$

Similarly,

$$\begin{aligned}\sigma_V^2 &= V(X\sin\alpha-Y\cos\alpha)=\sin^2\alpha.\sigma_1^2+\cos^2\alpha.\sigma_2^2\\ \text{Cov}(U,V) &= E[\{U-E(U)\}\{V-E(V)\}]\\ &= E\Big[\{(X-E(X))\cos\alpha+\{(Y-E(Y))\sin\alpha\}\\ &\qquad\times\{(X-E(X))\sin\alpha-(Y-E(Y))\cos\alpha\}\Big]\\ &=\sin\alpha\cos\alpha\,V(X)-\cos^2\alpha\,\text{Cov}(X,Y)\\ &\qquad+\sin^2\alpha\,\text{Cov}(X,Y)-\sin\alpha\cos\alpha\,V(Y)\\ &=(\sigma_1^2-\sigma_2^2)\sin\alpha\cos\alpha \qquad \text{[Using (1)]}\end{aligned}$$

Now $$\rho^2=\frac{[\text{Cov}(U,V)]^2}{\sigma_U^2\sigma_V^2}$$

where
$$\begin{aligned}\sigma_U^2\sigma_V^2 &=(\cos^2\alpha\,\sigma_1^2+\sin^2\alpha\,\sigma_2^2)(\sin^2\alpha\,\sigma_1^2+\cos^2\alpha\,\sigma_2^2)\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4)+\sigma_1^2\sigma_2^2(\cos^4\alpha+\sin^4\alpha)\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4)+\sigma_1^2\sigma_2^2[(\sin^2\alpha+\cos^2\alpha)^2-2\sin^2\alpha\cos^2\alpha]\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4-2\sigma_1^2\sigma_2^2)+\sigma_1^2\sigma_2^2\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^2-\sigma_2^2)^2+\sigma_1^2\sigma_2^2\end{aligned}$$

$$\therefore \quad \rho^2=\frac{(\sigma_1^2-\sigma_2^2)^2.\sin^2\alpha\cos^2\alpha}{\sigma_1^2\sigma_2^2+\sin^2\alpha\cos^2\alpha(\sigma_1^2-\sigma_2^2)^2}$$

$$=\frac{\frac{1}{4}(\sigma_1^2-\sigma_2^2)^2\sin^2 2\alpha}{\sigma_1^2\sigma_2^2+\sin^2 2\alpha.\,\frac{1}{4}(\sigma_1^2-\sigma_2^2)^2}$$

$$= \frac{(\sigma_1^2 - \sigma_2^2)^2}{4\sigma_1^2\sigma_2^2 \operatorname{cosec}^2 2\alpha + (\sigma_1^2 - \sigma_2^2)^2}$$

$$\Rightarrow \qquad \rho = \frac{\sigma_1^2 - \sigma_2^2}{[(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2 \operatorname{cosec}^2 2\alpha]^{1/2}}$$

Example 10·10. *If $U = aX + bY$ and $V = cX + dY$, where X and Y are measured from their respective means and if r is the correlation coefficient between X and Y, and if U and V are uncorrelated, show that*

$$\sigma_U \sigma_V = (ad - bc)\, \sigma_X \sigma_Y\, (1 - r^2)^{1/2}$$

[***Poona Univ. B.Sc., 1990; Delhi Univ. B.Sc. (Stat. Hons.), 1986***]

Solution. We have

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X\, \sigma_Y} \quad \Rightarrow \quad 1 - r^2 = 1 - \frac{[\text{Cov}(X, Y)]^2}{\sigma_X^2\, \sigma_Y^2}$$

$$\Rightarrow \qquad (1 - r^2)\, \sigma_X^2\, \sigma_Y^2 = \sigma_X^2\, \sigma_Y^2 - [\text{Cov}(X, Y)]^2 \qquad \ldots(*)$$

[This step is suggested by the answer]

$$U = aX + bY\,,\; V = cX + dY$$

Since X, Y are measured from their means,

$$\left.\begin{aligned} &E(X) = 0 = E(Y) \;\Rightarrow\; E(U) = 0 = E(V) \\ \text{and} \quad &\sigma_U^2 = E(U^2)\,;\; \sigma_V^2 = E(V^2) \end{aligned}\right\} \qquad \ldots(**)$$

Also $\quad aX + bY - U = 0 \quad$ and $\quad cX + dY - V = 0$

$$\Rightarrow \qquad \frac{X}{-bV + dU} = \frac{Y}{-cU + aV} = \frac{1}{ad - bc}$$

$$\Rightarrow \qquad \left.\begin{aligned} X &= \frac{1}{ad - bc}(dU - bV) \\ Y &= \frac{1}{ad - bc}(-cU + aV) \end{aligned}\right\} \qquad \ldots(***)$$

$$\therefore \qquad \text{Var}(X) = \frac{1}{(ad - bc)^2}[d^2 \sigma_U^2 + b^2 \sigma_V^2 - 2\, bd\, \text{Cov}(U, V)$$

$$= \frac{1}{(ad - bc)^2}[d^2 \sigma_U^2 + b^2 \sigma_V^2]$$

[Since U, V are uncorrelated $\Rightarrow$ Cov $(U, V) = 0$]

Similarly, we have

$$\text{Var}(Y) = \frac{1}{(ad - bc)^2}(c^2 \sigma_U^2 + a^2 \sigma_V^2)$$

$$\text{Cov}(X, Y) = E(XY) - E(X)\, E(Y) = E(XY) \qquad [\because E(X) = 0 = E(Y)]$$

$$= \frac{1}{(ad - bc)^2} E\big[(dU - bV)(-cU + aV)\big] \qquad [\text{From } (***)]$$

$$= \frac{1}{(ad-bc)^2} [-cd\,\sigma_U^2 - ab\,\sigma_V^2]$$

[Using (**) and Cov $(U, V) = 0$, given]

$$= \frac{-1}{(ad-bc)^2} [cd\,\sigma_U^2 + ab\,\sigma_V^2]$$

Substituting in (*), we get

$$(1-r^2)\,\sigma_X^2\,\sigma_Y^2 = \frac{1}{(ad-bc)^4} \times \left[(d^2\sigma_U^2 + b^2\sigma_V^2)(c^2\sigma_U^2 + a^2\sigma_V^2) - (cd\sigma_U^2 + ab\,\sigma_V^2)^2\right]$$

$$= \frac{1}{(ad-bc)^4} \times \left[c^2d^2\,\sigma_U^4 + a^2b^2\sigma_V^4 + (a^2d^2 + b^2c^2)\,\sigma_U^2\,\sigma_V^2 - c^2d^2\sigma_U^4 - a^2b^2\sigma_V^4 - 2abcd\,\sigma_U^2\,\sigma_V^2\right]$$

$$= \frac{1}{(ad-bc)^4}\left[a^2d^2 + b^2c^2 - 2abcd\right]\sigma_U^2\,\sigma_V^2$$

$$= \frac{1}{(ad-bc)^4}(ad-bc)^2\,\sigma_U^2\,\sigma_V^2$$

$$= \frac{\sigma_U^2\,\sigma_V^2}{(ad-bc)^2}$$

Cross multiplying and taking square root, we get the required result.

Example 10·11. (a) *Establish the formula :*

$$nr\sigma_X\sigma_Y = n_1r_1\sigma_{X_1}\sigma_{Y_1} + n_2r_2\sigma_{X_2}\sigma_{Y_2} + n_1dx_1dy_1 + n_2dx_2dy_2 \qquad ...(10·5)$$

where n_1, n_2 and n are respectively the sizes of the first, second and combined sample : $(\bar{x}_1, \bar{y}_1)$, $(\bar{x}_2, \bar{y}_2)$, $(\bar{x}, \bar{y})$, their means r_1, r_2 and r their coefficients of correlation; $(\sigma_{X_1}, \sigma_{Y_1})$, $(\sigma_{X_2}, \sigma_{Y_2})$, (σ_X, σ_Y) their standard deviations, and

$$dx_1 = \bar{x}_1 - \bar{x} \quad , \quad dy_1 = \bar{y}_1 - \bar{y}$$

$$dx_2 = \bar{x}_2 - \bar{x} \quad , \quad dy_2 = \bar{y}_2 - \bar{y}$$

(b) *Find the correlation co-efficient of combined sample given that*

	Sample I	*Sample II*
Sample size	*100*	*150*
Sample mean $(\bar{x})$	*80*	*72*
Sample mean $(\bar{y})$	*100*	*118*
Sample variance (σ_X^2)	*10*	*12*
Sample variance (σ_Y^2)	*15*	*18*
Correlation coefficient	*0·6*	*0·4*

Solution. (*a*) Let (x_{1i}, y_{1i}) ; $i = 1, 2, ..., n_1$ and (x_{2j}, y_{2j}); $j = 1, 2, ..., n_2$, be the two samples of sizes n_1 and n_2 respectively from the bivariate population. Then with the given notations, we have

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}, \quad \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

$$\left.\begin{aligned} n\sigma_X^2 &= n_1 (\sigma_{X_1}^2 + dx_1^2) + n_2 (\sigma_{X_2}^2 + dx_2^2) \\ n\sigma_Y^2 &= n_1 (\sigma_{Y_1}^2 + dy_1^2) + n_2 (\sigma_{Y_2}^2 + dy_2^2) \end{aligned}\right\} \quad \ldots(1)$$

$$r_1 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1)}{n_1\sigma_{X_1}\sigma_{Y_1}}, \quad r_2 = \frac{\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(y_{2j} - \bar{y}_2)}{n_2\sigma_{X_2}\sigma_{Y_2}} \quad \ldots(2)$$

For the pooled sample, we have

$$r = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x})(y_{1i} - \bar{y}) + \sum_{j=1}^{n_2} (x_{2j} - \bar{x})(y_{2j} - \bar{y})}{n\sigma_X \sigma_Y} \quad \ldots(3)$$

Now

$$\sum_{i=1}^{n_1} (x_{1i} - \bar{x})(y_{1i} - \bar{y}) = \sum_{i=1}^{n_1} \Big\{ \{(x_{1i} - \bar{x}_1) + (\bar{x}_1 - \bar{x})\} \{(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y})\} \Big\}$$

$$= \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y}) \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)$$

$$+ (\bar{x}_1 - \bar{x}) \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1) + n_1 (\bar{x}_1 - \bar{x})(\bar{y}_1 - \bar{y})$$

But $\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1) = 0$ and $\sum_{i=1}^{n_1} (y_{1i} - y_1) = 0,$

being the algebraic sum of the deviations from the mean.

$$\therefore \quad \sum_{i=1}^{n_1} (x_{1i} - \bar{x})(y_{1i} - \bar{y}) = n_1 r_1 \sigma_{X_1} \sigma_{Y_1} + n_1\, dx_1 dy_1 \qquad \text{[Using (2)]}$$

Similarly, we will get

$$\sum_{j=1}^{n_2} (x_{2j} - \bar{x})(y_{2j} - \bar{y}) = n_2 r_2 \sigma_{X_2} \sigma_{Y_2} + n_2\, dx_2 dy_2$$

Substituting in (3), we get the required formula.

(*b*) Here we are given :

$n_1 = 100$, $\bar{x}_1 = 80$, $\bar{y}_1 = 100$, $\sigma_{X_1}^2 = 10$, $\sigma_{Y_1}^2 = 15$, $r_1 = 0{\cdot}6$

$n_2 = 150$, $\bar{x}_2 = 72$, $\bar{y}_2 = 118$, $\sigma_{X_2}^2 = 12$, $\sigma_{Y_2}^2 = 18$, $r_2 = 0{\cdot}4$

$$\therefore \quad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{100 \times 80 + 150 \times 72}{100 + 150} = 75{\cdot}2$$

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{100 \times 100 + 150 \times 118}{100 + 150} = 110{\cdot}8$$

$$dx_1 = \bar{x}_1 - \bar{x} = 4{\cdot}8,\ dy_1 = \bar{y}_1 - \bar{y} = 10{\cdot}8$$

$$dx_2 = \bar{x}_2 - \bar{x} = 3{\cdot}2,\ dy_2 = \bar{y}_2 - \bar{y} = 7{\cdot}2$$

$$n\sigma_X^2 = n_1(\sigma_{X_1}^2 + dx_1^2) + n_2(\sigma_{X_2}^2 + dx_2^2) = 6640$$

$$n\sigma_Y^2 = n_1(\sigma_{Y_1}^2 + dy_1^2) + n_2(\sigma_{Y_2}^2 + dy_2^2) = 23640$$

Substituting these values in the formula and simplifying, we get

$$r = \frac{n_1 r_1 \sigma_{X_1}\sigma_{Y_1} + n_2 r_2 \sigma_{X_2}\sigma_{Y_2} + n_1 dx_1 dy_1 + n_2 dx_2 dy_2}{n\sigma_X\sigma_Y} = 0{\cdot}8186$$

Example 10·12. *The independent variables X and Y are defined by :*

$f(x) = 4ax,\ 0 \le x \le r$	$f(y) = 4by,\ 0 \le y \le s$
$= 0$, *otherwise*	$= 0$, *otherwise*

Show that :

$$Cov\,(U, V) = \frac{b-a}{b+a},$$

where $U = X + Y$ and $V = X - Y$

[I.I.T. (B. Tech.), Nov. 1992]

Solution. Since the total area under probability curve is unity (one), we have :

$$\int_0^r f(x)dx = 4a\int_0^r x\,dx = 1 \quad \Rightarrow \quad 2ar^2 = 1 \quad \Rightarrow \quad a = \frac{1}{2r^2} \qquad ...(i)$$

$$\int_0^r f(y)dy = 4b\int_0^s y\,dy = 1 \quad \Rightarrow \quad 2bs^2 = 1 \quad \Rightarrow \quad b = \frac{1}{2s^2} \qquad ...(ii)$$

$$\therefore\ f(x) = 4ax = \frac{2x}{r^2},\ 0 \le x \le r\ ;\ \text{and}\ f(y) = 4by = \frac{2y}{s^2},\ 0 \le y \le s \qquad ...(iii)$$

Since X and Y are independent variates,

$$r(X, Y) = 0 \quad \Rightarrow \quad \text{Cov}\,(X, Y) = 0 \qquad ...(iv)$$

$$\begin{aligned}
\text{Cov}\,(U, V) &= \text{Cov}\,(X + Y, X - Y)\\
&= \text{Cov}\,(X, X) - \text{Cov}\,(X, Y) + \text{Cov}\,(Y, X) - \text{Cov}\,(Y, Y)\\
&= \sigma_X^2 - \sigma_Y^2 && [\text{Using } (iv)]\\
\text{Var}\,(U) &= \text{Var}\,(X + Y) = \text{Var}\,(X) + \text{Var}\,(Y) + 2\,\text{Cov}\,(X, Y)\\
&= \sigma_X^2 + \sigma_Y^2 && [\text{Using } (iv)]\\
\text{Var}\,(Y) &= \text{Var}\,(X - Y) = \text{Var}\,(X) + \text{Var}\,(Y) - 2\,\text{Cov}\,(X, Y)\\
&= \sigma_X^2 + \sigma_Y^2 && [\text{Using } (iv)]
\end{aligned}$$

$$\therefore \quad r(U,V) = \frac{\text{Cov}(U,V)}{\sigma_U \sigma_V} = \frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2} \quad \ldots (v)$$

We have :

$$E(X) = \int_0^r x f(x)\, dx = \frac{2}{r^2}\int_0^r x^2\, dx = \frac{2r}{3} \quad \text{[From } (iii)]$$

$$E(X^2) = \int_0^r x^2 f(x)\, dx = \frac{2}{r^2}\int_0^r x^3 dx = \frac{r^2}{2}$$

$$\therefore \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{r^2}{2} - \frac{4r^2}{9} = \frac{r^2}{18} = \frac{1}{36a} \quad \text{[From } (i)]$$

Similarly, we shall get

$$E(Y) = \frac{2s}{3},\ E(Y^2) = \frac{s^2}{2} \text{ and Var}(Y) = \frac{s^2}{18} = \frac{1}{36b}$$

Substituting in (v), we get

$$r(U,V) = \frac{1/(36a) - 1/(36b)}{1/(36a) + 1/(36b)} = \frac{b-a}{b+a}.$$

Example 10·13. *Let the random variable X have the marginal density*

$$f_1(x) = 1, -\frac{1}{2} < x < \frac{1}{2}$$

and let the conditional density of Y be

$$\left.\begin{aligned} f(y \mid x) &= 1, x < y < x+1, -\frac{1}{2} < x < 0 \\ &= 1, -x < y < 1-x, 0 < x < \frac{1}{2} \end{aligned}\right\} \quad (*)$$

Show that the variables X and Y are uncorrelated.

Solution. We have

$$E(X) = \int_{-\frac{1}{2}}^{\frac{1}{2}} x f_1(x)\, dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x.1.dx = \left|\frac{x^2}{2}\right|_{-1/2}^{1/2} = 0$$

If $f(x, y)$ is the joint p.d.f. of X and Y, then

$$f(x,y) = f(y \mid x) f_1(x) = f(y \mid x). \quad (**) \quad [\because f_1(x) = 1]$$

$$E(XY) = \int_{-\frac{1}{2}}^{0}\int_{x}^{x+1} xy\, f(x,y)\, dx\, dy + \int_{0}^{\frac{1}{2}}\int_{-x}^{1-x} xy f(x,y)\, dx\, dy$$

$$= \int_{-\frac{1}{2}}^{0}\left[x\int_{x}^{x+1} y\, dy\right] dx + \int_{0}^{\frac{1}{2}}\left[x\int_{-x}^{1-x} y\, dy\right] dx \text{ [From } (*) \text{ and}$$

$$= \frac{1}{2} \int_{-\frac{1}{2}}^{0} x(2x+1)dx + \frac{1}{2} \int_{0}^{\frac{1}{2}} x(1-2x)\,dx$$

$$= \frac{1}{2}\left[\frac{2}{3}x^3 + \frac{x^2}{2}\right]_{-\frac{1}{2}}^{0} + \frac{1}{2}\left[\frac{x^2}{2} - \frac{2}{3}x^3\right]_{0}^{\frac{1}{2}}$$

$$= \frac{1}{2}\left[\frac{1}{12} - \frac{1}{8} - \frac{1}{12} + \frac{1}{8}\right] = 0$$

$\therefore \quad \text{Cov}(XY) = E(XY) - E(X)\,E(Y) = 0 \quad \Rightarrow \quad r(X, Y) = 0$

Hence the variables X and Y are uncorrelated.

EXERCISE 10(a)

1. (*a*) Show that the co-efficient of correlation r is independent of a change of scale and origin of the variables. Also prove that for two independent variables $r = 0$. Show by an example that the converse is not true. State the limits between which r lies and give its proof.

[Delhi Univ. M.Sc. (O.R.), 1986]

(*b*) Let ρ be the correlation coefficient between two jointly distributed random variables X and Y. Show that $|\rho| \le 1$ and that $|\rho| = 1$ if and only if X and Y are linearly related. *[Indan Forest Service, 1991]*

2. (*a*) Calculate the coefficient of correlation between X and Y for the following :

X...	1	3	4	5	7	8	10
Y...	2	6	8	10	14	16	20

Ans. $r(X, Y) = +1$

(*b*) Discuss the statistical validity of the following statements :

(*i*) "High positive coefficient of correlation between increase in the sale of newspapers and increase in the number of crimes leads to the conclusion that newspaper reading may be responsible for the increase in the number of crimes."

(*ii*) "A high positive value of r between the increase in cigarette smoking and increase in lung cancer establishes that cigarette smoking is responsible for lung cancer."

(*c*) (*i*) Do you agree with the statement that "$r = 0·8$ implies that 80% of the data are explained."

(*ii*) Comment on the following :

"The closeness of relationship between two variables is proportional to r".

Hint. (*a*) No (*b*) Wrong.

(*d*) By effecting suitable change of origin and scale, compute the product moment correlation coefficient for the following set of 5 observations on (X, Y) :

X :	−10	−5	0	5	10
Y :	5	9	7	11	13

Ans. $r(X, Y) = 0·34$

3. The marks obtained by 10 students in Mathematics and Statistics are given below. Find the coefficient of correlation between the two subjects.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics :	75	30	60	80	53	35	15	40	38	48
Marks in Statistics :	85	45	54	91	58	63	35	43	45	44

4. (*a*) The following table gives the number of blind per lakh of population in different age-groups. Find out the correlation between age and blindness.

Age in years :	0—10	10—20	20—30	30—40	40—50
Number of blind per lakh :	55	67	100	111	150
Age in year :	50—60	60—70	70—80		
Number of blind per lakh :	200	300	500		

Ans. 0·89

(*b*) The following table gives the distribution of items of production and also the relatively defective items among them, according to size-groups. Is there ay correlation between size and defect in quality ?

Size-Group :	15—16	16—17	17—18	18—19	19—20	20—21
No. of Items :	200	270	340	360	400	300
No. of defective items	150	162	170	180	180	120

Hint. Here we have to find the correlation coefficient between the size-group (X) and the percentage of defectives (Y) given below.

X	15·5	16·5	17·5	18·5	19·5	20·5
Y	75	60	50	50	45	40

Ans. $r = 0{\cdot}94$.

5. Using the formula

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\,r(X, Y)\,\sigma_X\,\sigma_Y$$

obtain the correlation coefficient between the heights of fathers (X) and of the sons (Y) from the following data :

X :	65	66	67	68	69	70	71	67
Y :	67	68	64	72	70	67	70	68

6. (*a*) From the following data, compute the co-efficient of correlation between X and Y.

	X series	*Y series*
No. of items	15	15
Arithmetic mean	25	18
Sum of squares of deviations from mean	136	138

Summation of product of deviations of X and Y series from the respective arithmetic means = 122.

Ans. $r(X, Y) = 0.891$

(*b*) Coefficient of correlation between two variables X and Y is 0·32. Their covariance is 7·86. The variance of X is 10. Find the standard deviation of Y series.

(*c*) In two sets of variables X and Y with 50 observations each, the following data were observed :

$$\bar{X} = 10, \sigma_X = 3, \bar{Y} = 6, \sigma_Y = 2 \text{ and } r(X, Y) = 0.3$$

But on subsequent verification it was found that one value of X (= 10) and one value of Y (= 6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of r affected ?

(Nagpur Univ. B.Sc., 1990)

Hint. $\Sigma X = n\bar{X} = 500, \Sigma Y = n\bar{Y} = 300$

$$\Sigma X^2 = n(\sigma_X^2 + \bar{X}^2) = 5450, \Sigma Y^2 = 50(4 + 36) = 2000$$

$$r\,\sigma_X \sigma_Y = \text{Cov}(X, Y) = \frac{\Sigma XY}{n} - \bar{X}\,\bar{Y}$$

$$\Rightarrow \quad 0.3 \times 3 \times 2 = \frac{\Sigma XY}{50} - 10 \times 6$$

$$\Rightarrow \quad \Sigma XY = 50(1.8 + 60) = 3090$$

After weeding out the incorrect pair of observation, *viz.*, ($X = 10, Y = 6$), the corrected values of $\Sigma X, \Sigma Y, \Sigma X^2, \Sigma Y^2$ and ΣXY for the remaining 50 − 1 = 49 pairs of observations are given below :

Corrected Values :

$\Sigma X = 500 - 10 = 490$; $\Sigma Y = 300 - 6 = 294$

$\Sigma XY = 3090 - 10 \times 6 = 3090 - 60 = 3030$

$\Sigma X^2 = 5450 - 10^2 = 5350$, $\Sigma Y^2 = 2000 - 6^2 = 1964$

$$\therefore r = \frac{\text{Corrected Cov}(X, Y)}{(\text{Corrected } \sigma_X) \times (\text{Corrected}) \, \sigma_Y} = \frac{90/49}{\sqrt{\frac{450}{49} \times \frac{200}{49}}} = 0.3$$

Hence the correlation coefficient is invariant in this case.

(*d*) A prognostic test in Mathematics was given to 10 students who were about to begin a course in Statistics. The scrores (X) in their test were examined in relations to scores (Y) in the final examination in Statistics. The following results were obtained :—

$$\Sigma X = 71, \Sigma Y = 70, \Sigma X^2 = 555, \Sigma Y^2 = 526 \text{ and } \Sigma XY = 527$$

Find the coefficient of correlation between X and Y.

(Kerala Univ. B.Sc., 1990)

7. (*a*) X_1 and X_2 are independent variables with means 5 and 10 and standard deviations 2 and 3 respectively. Obtain $r(U, V)$ where

$$U = 3X_1 + 4X_2 \text{ and } V = 3X_1 - X_2$$

Ans. 0 ***(Delhi Univ. B.Sc., 1988)***

(*b*) If X and Y are normal and independent with zero means and standard deviations 9 and 12 respectively, and if $X + 2Y$ and $kX - Y$ are non-correlated, find k.

(*c*) X, Y, Z are random variables each with expectation 10 and variances 1, 4 and 9 respectively. The correlation coefficients are

$$r(X, Y) = 0, r(Y, Z) = r\,(X, Y) = 1/4$$

Obtain the numerical values of :

(*i*) $E(X + Y - 2Z)$, (*ii*) Cov $(X + 3, Y + 3)$, (*iii*) $V(X - 2\,Z)$ and (*iv*) Cov $(3X, 5Z)$

Ans. (*i*) = 0, (*ii*) 0, (*iii*) 34, and (*iv*) 45/4.

(*d*) X and Y are discrete random variables. If Var (X) = Var$(Y) = \sigma^2$, Cov $(X, Y) = \frac{\sigma^2}{2}$, find (*i*) Var $(2X - 3Y)$, (*ii*) Corr $(2X + 3, 2Y - 3)$.

8\. (*a*) Prove that :

$$V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2ab \text{ Cov } (X, Y)$$

Hence deduce that if X and Y are independent

$$V(X \pm Y) = V(X) + V(Y)$$

(*b*) Prove that correlation coefficient between X and Y is positive or negative according as

$$\sigma_{X+Y} > \text{ or } < \sigma_{X-Y}$$

9\. Show that if X and Y are two random variables each assuming only two values and the correlation co-efficient between them is zero, then they are independent. Indicate with justification whether the result is true in general.

Find the correlation coefficient between X and $a - X$, where X is any random variable and a is constant.

10\. (*a*) X_i $(i = 1, 2, 3)$ are uncorrelated variables each having the same standard deviation. Obtain the correlation between $X_1 + X_2$ and $X_2 + X_3$.

Ans. 1/2

(*b*) If X_i $(i = 1, 2, 3)$ are three uncorrelated variables having standard deviations σ_1, σ_2 and σ_3 respectively, obtain the coefficient of correlation between $(X_1 + X_2)$ and $(X_2 + X_3)$.

Ans. $\sigma_2^2/\sqrt{(\sigma_1^2 + \sigma_2^2)\,(\sigma_2^2 + \sigma_3^2)}$

(*c*) Two random variables X and Y have zero means, the same variance σ^2 and zero correlation. Show that

$U = X \cos \alpha + Y \sin \alpha$ and $V = X \sin \alpha - Y \cos \alpha$

have the same variance σ^2 and zero correlation.

(Bangalore Univ. B.Sc., 1991)

(*d*) Let X and Y be uncorrelated random variables. If $U = X + Y$ and $V = X - Y$, prove that the coefficient of correlation between U and V is $(\sigma_X^2 - \sigma_Y^2)/(\sigma_X^2 + \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are variances of X and Y respectively.

(*e*) Two independent random variables X and Y have the following variances : $\sigma_X^2 = 36, \sigma_Y^2 = 16$. Calculate the coefficient of correlation between

$$U = X + Y \text{ and } V = X - Y$$

(f) Random variables X and Y have zero means and non-zero variances σ_X^2 and σ_Y^2. If $Z = Y - X$, then find σ_Z and the correlation coefficient $\rho(X, Z)$ of X and Z in terms of σ_X, σ_Y and the correlation coefficient $\rho\ (X, Y)$ of X and Y.

(g) If the independant random variables X_1, X_2 and X_3 have the means 4, 9 and 3 and variances 3, 7, 5, respectively, obtain the mean and variance of

(i) $Y = 2X_1 - 3X_2 + 4X_3$, (ii) $Z = X_1 + 2X_2 - X_3$, and

(iii) Calculate the correlation between Y and Z.

[*Delhi Univ. M.A.(Eco.)*., *1989*]

11. (a) $X_1, X_2, \ldots, X_n$ are uncorrelated random variables, all with the same distribution and zero means. Let $\bar{X} = \Sigma X_i / n$

Find the correlation co-efficient between (i) X_i and $\bar{X}$ and (ii) $X_i - \bar{X}$ and $\bar{X}$.

[*Delhi Univ. B.Sc. (Stat. Hons.)*, *1993*]

Hint.
$$r(X_i, \bar{X}) = \frac{\sigma^2/n}{\sqrt{\sigma^2 . \sigma^2/n}} = \frac{1}{\sqrt{n}}$$

$$\begin{aligned}\text{Cov}(X_i - \bar{X}, \bar{X}) &= \text{Cov}\ (X_i, \bar{X}) - \text{Var}\ (\bar{X}) \\ &= (\sigma^2/n) - (\sigma^2/n) = 0\end{aligned}$$

$$\therefore \qquad r(X_i - \bar{X}, \bar{X}) = 0$$

(b) $X_1, X_2, \ldots \ldots, X_n$ are random variables each with the same expected value μ and s.d. σ. The correlation coefficient between any two X's is ρ. Show that (i) $\text{Var}\ (\bar{X}) = \dfrac{\sigma^2}{n} + \left(1 - \dfrac{1}{n}\right)\rho\sigma^2$,

(ii) $E \sum\limits_1^n (X_i - \bar{X})^2 = (n-1)(1-\rho)\sigma^2$, and (iii) $\rho > -\dfrac{1}{n-1}$

12. (a) If X and Y are independent random variables, show that
$$r(X + Y, X - Y) = r^2(X, X + Y) - r^2\ (Y, X + Y),$$
where $r(X + Y, X - Y)$ denotes the co-efficient of correlation between $(X + Y)$ and $(X - Y)$. (*Meerut Univ. B.Sc., 1991*)

(b) Let X and Y be random variables having mean 0, variance 1 and correlation r. Show that $X - rY$ and Y are uncorrelated and that $X - rY$ has mean zero and variance $1 - r^2$.

13. X_1 and X_2 are two variables with zero means, variances σ_1^2 and σ_2^2 respectively and r is the correlation coefficient between them. Determine the values of the constants a and b which are independent of r such that $X_1 + aX_2$ and $X_1 + bX_2$ are uncorrelated.

14. (a) If X_1 and X_2 are two random variables with means μ_1 and μ_2, variances σ_1^2, σ_2^2 and correlation coefficient r, find the correlation co-efficient between

$$U = a_1 X_1 + a_2 X_2 \text{ and } V = b_1 X_1 + b_2 X_2,$$

where a_1, a_2 and b_1, b_2 are constants.

(*b*) Let X_1, X_2 be independent random variables with means μ_1, μ_2 and non-zero variances σ_1^2, σ_2^2 respectively. Let $U = X_1 - X_2$ and $V = X_1 X_2$. Find the correlation coefficient between (*i*) X_1 and U, (*ii*) X_1 and V, in terms of μ_1, μ_2, σ_1^2, σ_2^2.

15. (*a*) If $U = aX + bY$ and $V = bX - aY$, where X and Y are measured from their respective means and if U and V are uncorrelated, r the co-efficient of correlation between X and Y is given by the equation.

$$\sigma_U \sigma_V = (a^2 + b^2) \sigma_X \sigma_Y (1 - r^2)^{1/2}$$ **(Utkal Univ. B. Sc., 1993)**

(*b*) Let $U = aX + bY$ and $V = aX - bY$ where X, Y represent deviations from the means of two measurements on the same individual. The coefficient of correlation between X and Y is ρ. If U, V are uncorrelated, show that

$$\sigma_U \sigma_V = 2ab\sigma_X \sigma_Y (1 - r^2)^{1/2}$$

16. Show that, if a and b are constants and r is the correlation coefficient between X and Y, then the correlation coefficient between aX and bY is equal to r if the signs of a and b are alike, and to $-r$ if they are different.

Also show that, if constants a, b and c are positive, the correlation coefficient between $(aX + bY)$ and cY is equal to

$$(ar\sigma_X + b\sigma_Y) / \sqrt{(a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y)}$$

17. If X_1, X_2 and X_3 are three random variables measured from their respective means as origin and of equal variances, find the coefficient of correlation between $X_1 + X_2$ and $X_2 + X_3$ in terms of r_{12}, r_{13} and r_{23} and show that it is equal to

$$(i) \frac{r_{12} + 1}{2}, \text{ if } r_{13} = r_{23} = 0, \text{ and } (ii) = \frac{r_{12} + 3}{4}, \text{ if } r_{13} = r_{23} = 1$$

18. (*a*) For a weighted distribution (x_i, w_i), $(i = 1, 2, \ldots, n)$ show that the weighted arithmetic mean $\bar{x}_w = \Sigma w_i x_i / \Sigma w_i >$ or $<$ the unweighted mean $\bar{x} = \Sigma x_i / n$ according as $r_{xw} >$ or < 0.

(*b*) Given N values $x_1, x_2, \ldots, x_N$ of variable X and weights $w_1, w_2, \ldots, w_N$, express the coefficient of correlation between X and W in terms involving the difference between the arithmetic mean and the weighted mean of X.

19. (*a*) A coin is tossed n times. If X and Y denote the (random) number of heads and number of tails turned up respectively, show that $r(X, Y) = -1$.

Hint. Note that $X + Y = n \Rightarrow Y = n - X$

$$\therefore \quad r(X, Y) = r(X, n - X) = r(X, -X) = -r(X, X) = -1.$$

(*b*) Two dice are thrown, their scores being a and b. The first die is left on the table while the second is picked up and thrown again giving the score c. Suppose the process is repeated a large number of times. What is the correlation coefficient between $X = a + b$ and $Y = a + c$?

Ans. $r(X, Y) = \dfrac{1}{2}$

20. (*a*) If X and Y are independent random variables with means μ_1 and μ_2 and variances σ_1^2, σ_2^2 respectively, show that the correlation coefficient between $U = X$ and $V = X - Y$ in terms of μ_1, μ_2, σ_1^2 and σ_2^2 is $\sigma_1 / \sqrt{\sigma_1^2 + \sigma_2^2}$.

(*b*) If X and Y are independent random variables with non-zero variances, show that the correlation coefficient between $U = XY$ and $V = X$ in terms of mean and variance of X and Y is given by

$$\mu_2\sigma_1 / \sqrt{\sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}$$

[*Delhi Univ. B.Sc. (Stat Hons.), 1987*]

21. If X_i, Y_j and Z_k are all independent random variables with mean zero and unit variance, find the correlation coefficient between

$$U = \sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j \text{ and } V = \sum_{i=1}^{m} X_i + \sum_{k=1}^{n} Z_k$$

Ans. $r(U, V) = m/(m+n)$ (*Bombay Univ., B.Sc, 1990*)

22. (*a*) Find the value of l so that the correlation coefficient between $(X - lY)$ and $(X + Y)$ is maximum, where X, Y are independent random variables each with mean zero and variance 1. [**Ans.** $l = -1$]

Hint. $U = X - lY$; $V = X + Y$. Now find l so that $r(U, V) = 1$.

(*b*) If $U = X + kY$ and $V = X + mY$ and r is the correlation coefficient between X and Y, find the correlation coefficient between U and V. Show that U and V are uncorrelated if

$$k = \frac{-\sigma_X(\sigma_X + rm\,\sigma_Y)}{\sigma_Y(r\,\sigma_X + m\sigma_Y)}$$

and further if $m = \dfrac{\sigma_X}{\sigma_Y}$, then $k = -\dfrac{\sigma_X}{\sigma_Y}$. (*Gujarat Univ. M.A., 1993*)

23. X_1, X_2, X_3 are three variables, each with variance σ^2 and the correlation coefficient between any two of them is r. If $\bar{X} = (X_1 + X_2 + X_3)/3$, show that

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{3}(1 + 2r)$$

Deduce that $r \geq -1/2$.

24. (*a*) If $U = aX + bY$ and $V = bX - aY$, show that U and V are uncorrelated if

$$\frac{ab}{a^2 - b^2} = \frac{\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}$$

where ρ is the coefficient of correlation between X and Y. Show further that, in this case

$$\sigma_U^2 + \sigma_V^2 = (a^2 + b^2)(\sigma_X^2 + \sigma_Y^2) \text{ and } \sigma_U\sigma_V = (a^2 + b^2)\sigma_X\sigma_Y\sqrt{1 - \rho^2}$$

(*b*) If $u = aX + bY$, $v = cX + dY$, show that

$$\begin{vmatrix} \text{var}(u) & \text{cov}(u, v) \\ \text{cov}(u, v) & \text{var}(v) \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix}^2 \begin{vmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \text{var}(Y) \end{vmatrix}$$

25. If X is a standard normal variate and $Y = a + bX + cX^2$, where a, b, c are constants, find the correlation coefficient between X and Y. Hence or otherwise obtain the conditions when (*i*) X and Y are uncorrelated and (*ii*) X and Y are perfectly correlated.

26. (*a*) If $X \sim N(0, 1)$, find corr (X, Y) where $Y = a + bX + cX^2$.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1985*]

Ans. $r(X, Y) = \dfrac{b}{\sqrt{b^2 + 2c^2}}$

(*b*) If X has Laplace distribution with parameters $(\lambda, 0)$ and $Y = a + bX + cX^2$, find $\rho(X, Y)$

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989*]

Hint. $p(x) = \frac{1}{2}\lambda \exp[-\lambda\,|x|], \ -\infty < x < \infty.$

$$E(X^{2k+1}) = 0 = \mu_{2k+1}\ ;\ E(X^{2k}) = \mu_{2k} = (2k)!\,/\,\lambda^{2k}$$

$$\rho_{XY} = \frac{\lambda b}{\sqrt{b^2\lambda^2 + 10c^2}}$$

27. In a sample of n random observations from exponential distribution with parameter λ, the number of observations in $(0, 1/\lambda)$ and $(1/\lambda, 2/\lambda)$, denoted by X and Y are noted. Find $\rho(X, Y)$.

Hint. $p_1 = p(0 < X < 1/\lambda) = \displaystyle\int_0^{1/\lambda} \lambda e^{-\lambda x} dx = \frac{e-1}{e}$

$$p_2 = p(1/\lambda \le Y < 2/\lambda) = \int_{1/\lambda}^{2/\lambda} \lambda e^{-\lambda y} dy = \frac{e-1}{e^2}$$

Then (X, Y) has a trinomial distribution with parameters $(n = 3, p_1, p_2, p_3 = 1 - p_1 - p_2)$.

Hence we have

$$\rho(X, Y) = -\left[\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}\right]^{1/2} = -\frac{e-1}{\sqrt{e^2 - e + 1}}.$$

28. Prove that :

$$r(X, Y + Z) = \frac{\sigma_Y}{\sigma_{Y+Z}}\cdot r(X, Y) + \frac{\sigma_Z}{\sigma_{Y+Z}}\cdot r(X, Z)$$

29. If X and Y are independent random variables, find Corr(X, XY). Deduce the value of Corr$(X, X/Y)$.

Ans. $r(X, XY) = \sigma_X\,\mu_Y/[\sigma_X^2\sigma_Y^2 + \mu_X^2\,\sigma_Y^2 + \mu_Y^2\,\sigma_X^2]^{1/2}$

30. Prove or Disprove :

(*a*) $r(X, Y) = 0 \quad \Rightarrow \quad r(|X|, Y) = 0$

(*b*) $r(X, Y) = 0,\ r(Y, Z) = 0 \ \Rightarrow\ r(X, Z) = 0.$

Ans. (*a*) False, unless X and Y are independent.

(*b*) **Hint.** Let $Z \equiv X$, and X and Y be independent. Then $r(X, Y) = 0 = r(Y, Z)$. But $r(X, Z) = r(X, X) = 1$.

31. Let random variable X have a p.d.f. $f(.)$ with distribution function $F(.)$, mean μ and variance σ^2. Define $Y = \alpha + \beta X$, where α and β are constants satisfying $-\infty < \alpha < \infty$, and $\beta > 0$.

(*a*) Select α and β so that Y has mean 0 and variance 1.

(*b*) What is the correlation coefficient between X and Y?

32. Let (X, Y) be jointly discrete random variables such that each X and Y have at most two mass points. Prove or disprove : X and Y are independent if and only if they are uncorrelated.

Ans. True.

33. If the variables $X_1, X_2, \ldots, X_{2n}$ all have the same variance σ^2 and the correlation coefficient between X_i and X_j $(i \neq j)$ has the same value, show that the correlation between $\sum_{i=1}^{n} X_i$ and $\sum_{j=n+1}^{2n} X_j$ is given by $[n\rho/\{1 + (n-1)\rho\}]$.

34. The means of independent r.v's $X_1, X_2, \ldots, X_n$ are zero and variances are equal, say unity. The correlation coefficients between the sum of selected t $(< n)$ variables out of these variables and the sum of all n variables are found out. Prove that the sum of squares of all these correlation coefficients is ${}^{n-1}C_{t-1}$.

[*Burdwan Univ. B.Sc. (Hons.), 1989*]

35. Two variables U and V are made up of the sum of a number of terms as follows :

$$U = X_1 + X_2 + \ldots + X_n + Y_1 + Y_2 + \ldots + Y_a ,$$
$$V = X_1 + X_2 + \ldots + X_n + Z_1 + Z_2 + \ldots + Z_b ,$$

where a and b are all suffixes and where X's, Y's and Z's are all uncorrelated standardised random variables. Show that the correlation coefficient between U and V is $\dfrac{n}{\sqrt{(n+a)(n+b)}}$. Show further that

$$\left.\begin{aligned} \xi &= \sqrt{(n+b)}\, U + \sqrt{(n+a)}\, V \\ \text{and} \qquad \eta &= \sqrt{(n+b)}\, U - \sqrt{(n+a)}\, V \end{aligned}\right\} \qquad \ldots(*)$$

are uncorrelated [*South Gujarat Univ. B.Sc., 1989*]

36. (*a*) Let the random variables X and Y have the joint p.d.f.

$$f(x, y) = 1/3 \; ; \; (x, y) = (0, 0), (1, 1) \; (2, 0)$$

Compute $E(X)$, $V(X)$, $E(Y)$, $V(Y)$ and $r(X, Y)$. Are X and Y stochastically independent ? Give reasons.

(*b*) Let (X, Y) have the probability distribution :

$$f(0, 0) = 0{\cdot}45, f(0, 1) = 0{\cdot}05, f(1, 0) = 0{\cdot}35, f(1, 1) = 0{\cdot}15.$$

Evaluate $V(X)$, $V(Y)$ and $\rho(X, Y)$.

Show that while X and Y are correlated, X and $X-5Y$ are uncorrelated. Are X and $X - 5Y$ independent ?

(*c*) Given the bivariate probability distribution :

$$\begin{array}{lll} f(-1, 0) = 1/15, & f(-1, 1) = 3/15, & f(-1, 2) = 2/15 \\ f(0, 0) = 2/15, & f(0, 1) = 2/15, & f(0, 2) = 1/15 \\ f(1, 0) = 1/15, & f(1, 1) = 1/15, & f(1, 2) = 2/15 \end{array}$$
$$f(x, y) = 0, \text{ elsewhere.}$$

Obtain :

(*i*) The marginal distributions of X and Y.

(ii) The conditional distributions of Y given $X = 0$.

(iii) $E(Y|X = 0)$.

(iv) The product moment correlation coefficient between X and Y.
Are X and Y independently distributed ?

37. If X and Y are standardised variates with correlation coefficient ρ, prove that
$$E\,[\max\,(X^2, Y^2)] \le 1 + \sqrt{1-\rho^2}$$

Hint. $\max\,(X^2, Y^2) = \frac{1}{2}|X^2 - Y^2| + \frac{1}{2}(X^2 + Y^2)$ ...(*)
$$E(X) = E(Y) = 0;\ E(X^2) = E(Y^2) = 1;\ E(XY) = \rho$$
and $$[E\,|X - Y|\,.\,|X + Y|]^2 \le E\,(X - Y)^2\,.\,E\,(X + Y)^2$$
(By Cauchy-Schwartz Inequality)

38. The joint p.d.f. of two variates X and Y is given by
$$f(x, y) = k[(x + y) - (x^2 + y^2)]\ ;\ 0 < (x, y) < 1$$
$$= 0, \text{ otherwise.}$$
Show that X and Y are uncorrelated but not independent.

39(a). If the random variables X and Y have the joint p.d.f.,
$$f(x, y) = \begin{cases} x + y;\ 0 < x < 1,\ 0 < y < 1 \\ 0, \text{ elsewhere} \end{cases}$$
then show that the correlation coefficient between X and Y is $-\frac{1}{11}$.

[*Madras Univ. B.Sc., Oct., 1990*]

(b) The density function f of a random variable X is given by
$$f(x) = \begin{cases} kx^2, \text{ if } -1 \le x \le 1 \\ 0, \quad \text{otherwise} \end{cases}$$

(i) What is the value of k ? What is the distribution function of X ?

(ii) Obtain the density function of the random variable $Y = X^2$.

(iii) Obtain the correlation coefficient between X and Y.

(iv) Are X and Y independently distributed ?

40(a). If $f(x, y) = \dfrac{6 - x - y}{8}$; $0 \le x \le 2,\ 2 \le y \le 4,$
find (i) Var (X), (ii) Var (Y) (iii) $r\,(X, Y)$.

Ans. (i) $\frac{11}{36}$, (ii) $\frac{11}{36}$, (iii) $-\frac{1}{11}$.

(b) Given the joint density of random variables X, Y, Z as :
$$f(x, y, z) = k\,x \exp\,[-(y + z)],\ 0 < x < 2,\ y \ge 0,\ z \ge 0$$
$$= 0, \text{ elsewhere}$$
Find

(i) k,

(ii) the marginal density function,

(iii) conditional expectation of Y, given X and Z, and

(iv) the product moment correlation between X and Y.

[*Madras Univ. B.Sc. (Main Stat.), 1988*]

(c) Suppose that the two dimensional random variable (X, Y) has p.d.f. given by $\quad f(x, y) = ke^{-y}, 0 < x < y < 1$

$= 0$, elsewhere

Find the correlation coefficient r_{XY}. [*Delhi Univ. M.C.A., 1991*]

41. The joint density of (X, Y) is :

$$f(x, y) = \frac{1}{8}(x + y), \quad 0 \le x \le 2, 0 \le y \le 2.$$

Find $\mu'_{rs} = E(X^r Y^s)$ and hence find Corr (X, Y).

Ans. $\mu'_{rs} = 2^{r+s}\left[\dfrac{1}{(r+2)(s+1)} + \dfrac{1}{(r+1)(s+2)}\right]; r = -\dfrac{1}{11}$.

(b) Find the m.g.f. of the bivariate distribution :

$$f(x, y) = 1, 0 < (x, y) < 1$$
$$= 0, \text{ otherwise}$$

and hence find $r(X, Y)$.

Ans. $M(t_1, t_2) = (e^{t_1} - 1)(e^{t_2} - 1)/(t_1 t_2)\ ;\ t_1 \ne 0, t_2 \ne 0, r(X, Y) = 0$.

42. Let (X, Y) have joint density :

$$f(x, y) = e^{-(x+y)} I_{(0, \infty)}(x) \cdot I_{(0, \infty)}(y)$$

Find Corr (X, Y). Are X and Y independent ?

Ans. Corr $(X, Y) = 0$: X and Y are independent.

43. A bivariate distribution in two discrete random variables X and Y is defined by the probability generating function :

$$\exp[a(u-1) + b(v-1) + c(u-1)(v-1)],$$

simultaneous probability of $X = r \cap Y = s$, where r and s are integers being the coefficient of $u^r v^s$. Find the correlation coefficient between X and Y.

Hint. Put $u = e^{t_1}$ and $v = e^{t_2}$ in $\exp[a(u-1) + b(v-1) + c(u-1)(v-1)]$, the result will be the m.g.f. of a bivariate distribution and is given by

$$M(t_1, t_2) = \exp[a(e^{t_1} - 1) + b(e^{t_2} - 1) + c(e^{t_1} - 1)(e^{t_2} - 1)]$$

We have $\left[\dfrac{\partial M}{\partial t_1}\right]_{t_1 = t_2 = 0} = a, \quad \left[\dfrac{\partial^2 M}{\partial t_1^2}\right]_{t_1 = t_2 = 0} = a(a+1)$

$$\left[\frac{\partial^2 M}{\partial t_1 \partial t_2}\right]_{\substack{t_1 = 0\\ t_2 = 0}} = ab + c, \quad \left[\frac{\partial M}{\partial t_2}\right]_{\substack{t_1 = 0\\ t_2 = 0}} = b, \quad \left[\frac{\partial^2 M}{\partial t_2^2}\right]_{\substack{t_1 = 0\\ t_2 = 0}} = b(b+1)$$

So we have

$E(X) = a, E(X^2) = a(a+1), E(Y) = b, \ E(Y^2) = b(b+1)$ and $E(XY) = ab + c$

$$\therefore \quad r(X, Y) = \frac{E(XY) - E(X)\,E(Y)}{\sqrt{[E(X^2) - \{E(X)\}^2][E(Y^2) - \{E(Y)\}^2]}} = \frac{c}{\sqrt{ab}}$$

44. Let the number X be chosen at random from among the integers 1, 2, 3, 4 and the number Y be chosen from among those at least as large as X. Prove that Cov $(X, Y) = 5/8$. Find also the regression line of Y on X.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1990*]

Hint. $\quad P(X = k) = \dfrac{1}{4}\ ;\ k = 1, 2, 3, 4$ and $Y \ge X$.

$P(Y = y \mid X = 1) = \frac{1}{4}$; $y = 1, 2, 3, 4$ $(\because y \geq x)$;

$P(Y = y \mid X = 2) = \frac{1}{3}, y = 2, 3, 4$

$P(Y = y \mid X = 3) = \frac{1}{2}, y = 3, 4$; $P(Y = y \mid X = 4) = 1, y = 4.$

The joint probability distribution can be obtained on using :

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y \mid X = x).$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{5/8}{\sqrt{(5/4) \times (41/48)}} = \sqrt{\frac{15}{41}}$$

Regression line of Y on X : $Y - E(Y) = \frac{r \sigma_Y}{\sigma_X} [X - E(X)]$

45. Two ideal dice are thrown. Let X_1 be the score on the first dice ad X_2, the score on the second dice. Let $Y = \max \{X_1, X_2\}$. Obtain the joint distribution of Y and X_1 and show that

$$\text{Corr}((Y, X_1) = \frac{3}{2\sqrt{73}}$$

46. Consider an experiment of tossing two tetrahedra. Let X be the number of the down turned face of first tetrahedron and Y, the larger of the two numbers. Obtain the joint distribution of X and Y and hence $\rho(X, Y)$.

Ans. $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_y} = \frac{5/8}{\sqrt{5/4} . \sqrt{55/64}} = \frac{2}{\sqrt{1.1}}$

47. Three fair coins are tossed. Let X denote the number of heads on the first two coins and let Y denote the number of tails on the last two coins.

(*a*) Find the joint distribution of X and Y.

(*b*) Find the conditional distribution of Y given that $X = 1$.

(*c*) Find Cov. (X, Y)

Ans. Cov. $(X, Y) = -1/4$.

48. For the trinomial distribution of two random variables X and Y :

$$f(x, y) = \frac{n!}{x!\, y!\, (n - x - y)!} p^x q^y (1 - p - q)^{n - x - y}$$

for $x, y = 0, 1, 2, \ldots, n$ and $x + y \leq n$, $p \geq 0$, $q \geq 0$ and $p + q \leq 1$.

(*a*) Obtain the marginal distribution of Y

(*b*) Obtain $E(X \mid Y = y)$.

(*c*) Find $\rho(X, Y)$.

Ans. (*a*) $X \sim B(n, p)$, $Y \sim B(n, q)$

(*b*) $(X \mid Y = y) \sim B\left(n - y, \frac{p}{1 - q}\right)$

(Note : $p + q \neq 1$)

$$\therefore E(X \mid Y = y) = (n - y)\left(\frac{p}{1 - q}\right)$$

(c) $\text{Cov}(X, Y) = -npq$; $\rho(X, Y) = -\left[\dfrac{pq}{(1-p)(1-q)}\right]^{1/2}$

OBJECTIVE TYPE QUESTIONS

I. Comment on the following :

(i) $r_{XY} = 0 \Rightarrow X$ and Y are independent.

(ii) If $r_{XY} > 0$ then $r_{X,-Y} > 0$, $r_{-X,Y} > 0$ and $r_{-X,-Y} > 0$

(iii) $r_{XY} > 0 \Rightarrow E(XY) > E(X)\,E(Y)$

(iv) Pearson's coefficient of correlation is independent of origin but not of scale.

(v) The numerical value of product moment correlation coefficient 'r' between two variables X and Y cannot exceed unity.

(vi) If the correlation coefficient between the variables X and Y is zero then the correlation coefficient between X^2 and Y^2 is also zero.

(vii) If $r > 0$, then as X increases, Y also increases.

(viii) "The closeness of relationship between two variables is proportional to r."

(ix) r measures every type of relationship between the two variables.

II. Comment on the following values of 'r' (correlation coefficient) :

$$1,\ -0{\cdot}95,\ 0,\ -1{\cdot}64,\ 0{\cdot}87,\ 0{\cdot}32,\ -1,\ 2{\cdot}4.$$

III. (i) If $\rho_{XY} = -0{\cdot}9$, then for large values of X, what sort of values do we expect for Y ?

(ii) If $\rho_{XY} = 0$, what is the value of cov (X, Y) and how are X and Y related ?

IV. Indicate the correct answer :

(i) The coefficient of correlation will have positive sign when
(a) X is increasing, Y is decreasing, (b) both X and Y are increasing, (c) X is decreasing, Y is increasing, (d) there is no change in X and Y.

(ii) The coefficient of correlation (a) can take any value between -1 and $+1$ (b) is always less than -1, (c) is always more than $+1$, (d) cannot be zero.

(iii) The coefficient of correlation (a) cannot be positive, (b) cannot be negative, (c) is always positive, (d) can be both positive as well as negative.

(iv) Probable error of r is

(a) $0{\cdot}6475\dfrac{1-r^2}{\sqrt{n}}$, (b) $0{\cdot}6754\dfrac{1+r^2}{\sqrt{n}}$, (c) $0{\cdot}6547\dfrac{1-r^2}{n}$,

(d) $0{\cdot}6754\dfrac{1-r^2}{n}$.

(v) The coefficient of correlation between X and Y is 0·6. Their covariance is 4·8. The variance of X is 9. Then the S.D. of Y is

(a) $\dfrac{4{\cdot}8}{3 \times 0{\cdot}6}$, (b) $\dfrac{0{\cdot}6}{4{\cdot}8 \times 3}$, (c) $\dfrac{3}{4{\cdot}8 \times 0{\cdot}6}$, (d) $\dfrac{4{\cdot}8}{9 \times 0{\cdot}6}$.

(*vi*) The coefficient of correlation is independent of (*a*) change of scale only, (*b*) change of origin only, (*c*) both change of scale and origin, (*d*) neither change of scale nor change of origin.

V. Fill in the blanks :

(*i*) The Karl Pearson coefficient of correlation between variables X and Y is ... ...

(*ii*) Two independent variables are ... ...

(*iii*) Limits for correlation coefficient are ... ...

(*iv*) If r be the correlation coefficient between the random variables X and Y then the variance of $X + Y$ is ... ...

(*v*) The absolute value of the product moment correlation coefficient is less than ... ...

(*vi*) Correlation coefficient is i3nvariant under changes of .. and ... ...

VI. How can you use scatter diagram to obtain an idea of extent and nature (direction) of the correlation coefficient ?

10·4. Calculation of the Correlation Coefficient for a Bivariate Frequency Distribution. When the data are considerably large, they may be summarised by using a two-way table. Here, for each variable a suitable number of classes are taken, keeping in view the same considerations as in the univariate case. If there are n classes for X and m classes for Y, there will be in all $m \times n$ cells in the two-way table. By going through the pairs of values of X and Y, we can find the frequency for each cell. The whole set of cell frequencies will then define a *bivariate frequency distribution*. The column totals and row totals will give us the marginal distributions of X and Y. A particular column or row will be called the conditional distribution of Y for given X or of X for given Y respectively.

Suppose that the bivariate data on X and Y are presented in a two-way correlation table (shown on page 10·33) where there are m classes of Y placed along the horizontal line and n classes of X along a vertical line and f_{ij} is the frequency of individuals lying in the (i, j)th cell.

Here
$$\sum_x f(x, y) = g(y)$$
is the sum of the frequencies along any row and
$$\sum_y f(x, y) = f(x)$$
is the sum of the frequencies along any column. We observe that

$$\sum_x \sum_y f(x, y) = \sum_y \sum_x f(x, y) = \sum_x f(x) = \sum_y g(y) = N$$

Thus
$$\bar{x} = \frac{1}{N} \sum_x \sum_y x f(x, y) = \frac{1}{N}\left[\sum_x \left\{ x \sum_y f(x, y) \right\}\right] = \frac{1}{N} \sum_x x f(x)$$

Similarly
$$\bar{y} = \frac{1}{N} \sum_x \sum_y y f(x, y) = \frac{1}{N} \sum_y y \, g(y)$$

$$\sigma_X^2 = \frac{1}{N} \sum_x \sum_y x^2 f(x, y) - \bar{x}^2 = \frac{1}{N} \sum_x x^2 f(x) - \bar{x}^2$$

BIVARIATE FREQUENCY TABLE (CORRELATION TABLE)

Y Series ↓ / X Series →	Classes Mid Points x_1 x_2 ... x_i ... ... x_m	Total of frequencies of Y $g(y)$
y_1 y_2 ⋮ y_j ⋮ ⋮ y_n	$f(x, y)$	$g(y) = \sum_x f(x, y)$
Total of frequencies of X $f(x)$	$f(x) = \sum_y f(x, y)$	$N \rightarrow \sum_x \sum_y f(x, y)$ ↓ $\sum_y \sum_x f(x, y)$

Example 10·14. *The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test.*

Marks ↓ / Ages in years →	18	19	20	21	Total
10—20	4	2	2	—	8
20—30	5	4	6	4	19
30—40	6	8	10	11	35
40—50	4	4	6	8	22
50—60	—	2	4	4	10
60—70	—	2	3	1	6
Total	19	22	31	28	100

Calculate the correlation coefficient.

Solution.

CORRELATION TABLE

v	y	u → / x → / Marks ↓	-1 18	0 19	1 20	2 21	Total $f(v)$	$vf(v)$	$v^2f(v)$	$\sum\limits_u uv\, f(u, v)$
-2	15	10—20	8 4	0 2	-4 2		8	-16	32	4
-1	25	20—30	5 5	0 4	-6 6	-8 4	10	-19	19	-9
0	35	30—40	0 6	0 8	0 10	0 11	35	0	0	0
1	45	40—50	-4 4	0 4	6 6	16 8	22	22	22	18
2	55	50—60		0 2	8 4	16 4	10	20	40	24
3	65	60—70		0 2	9 3	6 1	6	18	54	15
		Total $f(u)$	19	22	31	28	100	25	167	52
		$uf(u)$	-19	0	31	56	68			
		$u^2f(u)$	19	0	31	112	162			
		$u\sum\limits_v vf(u, v)$	9	0	13	30	52			

Let $\qquad U = X - 19, V = \{(Y - 35)/10\}$

$$\bar{u} = \frac{1}{N} \sum_u u f(u) = \frac{68}{100} = 0.68, \quad \bar{v} = \frac{1}{N} \sum_v v\, g(v) = \frac{25}{100} = 0.25$$

$$\text{Cov}(u, v) = \frac{1}{N} \sum_u \sum_v uv\, f(u, v) - \bar{u}\,\bar{v} = \frac{1}{100} \times 52 - 0.68 \times 0.25 = 0.35$$

$$\sigma_U^2 = \frac{1}{N} \sum_u u^2 f(u) - \bar{u}^2 = \frac{162}{100} - (0.68)^2 = 1.1576$$

$$\sigma_V^2 = \frac{1}{N} \sum_v v^2 g(v) - \bar{v}^2 = \frac{167}{100} - (0.25)^2 = 1.6075$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U\, \sigma_V} = \frac{0.35}{\sqrt{1.1576 \times 1.6075}} = 0.25$$

Since correlation coefficient is independent of change of origin and scale,

$$r(X, Y) = r(U, V) = 0{\cdot}25$$

Remark. Figures in circles in the table on page10·34 are the product terms $uvf(u, v)$:

Example 10·15. *The joint probability distribution of X and Y is given below :*

Y \ X	−1	+1
0	$\frac{1}{8}$	$\frac{3}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$

Find the correlation coefficient between X and Y.

Solution.

COMPUTATION OF MARGINAL PROBABILITIES

Y \ X	−1	+1	$g(y)$
0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{4}{8}$
$p(x)$	$\frac{3}{8}$	$\frac{5}{8}$	1

We have :

$$E(X) = \Sigma\, xp(x) = (-1) \times \frac{3}{8} + 1 \times \frac{5}{8} = \frac{1}{4}$$

$$E(X^2) = \Sigma\, x^2 p(x) = (-1)^2 \times \frac{3}{8} + 1^2 \times \frac{5}{8} = 1$$

$$\therefore \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = 1 - \frac{1}{16} = \frac{15}{16}$$

$$E(Y) = \Sigma\, y\, g(y) = 0 \times \frac{4}{8} + 1 \times \frac{4}{8} = \frac{1}{2}$$

$$E(Y^2) = \Sigma\, y^2\, g(y) = 0^2 \times \frac{4}{8} + 1^2 \times \frac{4}{8} = \frac{1}{2}$$

$$\therefore \quad \text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$E(XY) = 0 \times (-1) \times \frac{1}{8} + 0 \times 1 \times \frac{3}{8} + 1 \times (-1) \times \frac{2}{8} + 1 \times 1 \times \frac{2}{8}$$

$$= -\frac{2}{8} + \frac{2}{8} = 0$$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\, E(Y) = 0 - \frac{1}{4} \times \frac{1}{2} = -\frac{1}{8}$$

$$\therefore \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-\frac{1}{8}}{\sqrt{\frac{15}{16} \times \frac{1}{4}}} = \frac{-1}{\sqrt{15}} = \frac{-1}{3{\cdot}873}$$

$$= -0{\cdot}2582$$

EXERCISE 10(b)

1. Write a brief note on the correlation table :

The following are the marks obtained by 24 students in a class test of Statistics and Mathematics :

Role No. of Students :	1	2	3	4	5	6	7	8	9	10	11	12
Marks in Statistics :	15	0	1	3	16	2	18	5	4	17	6	19
Marks in Mathematics :	13	1	2	7	8	9	12	9	17	16	6	18
Roll No. of Students :	13	14	15	16	17	18	19	20	21	22	23	24
Marks in Statistics :	14	9	8	13	10	13	11	11	12	18	9	7
Marks in Mathematics :	11	3	5	4	10	11	14	7	18	15	15	3

Prepare a correlation table taking the magnitude of each class interval as four marks and the first class interval as "equal to 0 and less than 4". Calculate Karl Pearson's coefficient of correlation between the marks in Statistics and marks in Mathematics from the correlation table.

Ans. 0·5544.

2. An employment bureau asked applicants their weekly wages on jobs last held. The actual wages were obtained for 54 of them; and are recorded in the table below; x represents reported wage, y actual wage, and the entry in the table represents frequency. Find the correlation coefficient and comment on the significance of the computed value. [Four figure log table may be used].

$y \rightarrow$ / $\downarrow x$	15	20	25	30	35	40
40						2
35				3	5	
30			4	15		
25			20			
20		3	1			
15	1					

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

3. Calculate the correlation coefficient from the following table :—

$x \rightarrow$ / $y \downarrow$	0—10	10—20	20—30	30—40
0—5	1	3	2	0
5—10	7	10	8	1
10—15	10	13	10	8
15—20	5	8	10	7
20—25	0	1	5	4

4. (*a*) Find the correlation coefficient between age and salary of 50 workers in a factory :

Age (in years) ↓	Daily pay in rupees				
	160—169	170—179	180—189	190—199	200—209
20—30	5	3	1	...	...
30—40	2	6	2	1	...
40—50	1	2	4	2	2
50—60	...	1	3	6	2
60—70	...	...	1	1	5

(*b*) Fnd the coefficient of correlation between the ages of 100 mothers and daughters :

Age of mothers in years (X)	Age of daughters in years (Y)					Total
	5—10	10—15	15—20	20—25	25—30	
15—25	6	3				9
25—35	3	16	10			29
35—45		10	15	7		32
45—55			7	10	4	21
55—65				4	5	9
Total	9	29	32	21	9	100

[*Madras Univ. B.Sc. (Main Maths.), 1991*]

5. Given the following frequency distribution of (X, Y) :

$X \rightarrow$ / $Y \downarrow$	5	10	Total
10	30	20	50
20	20	30	50
Total	50	50	100

find the frequency distribution of (U, V), where

$$U = \frac{X - 7\cdot5}{2\cdot5}, \quad V = \frac{Y - 15}{-5}.$$

What shall be the relationship between the correlation coefficients between X, Y, and U, V ?

6. (*a*) Find the coefficient of correlation between X and Y for the following table :

$X \downarrow$ \ $Y \rightarrow$	y_1	y_2	Total
x_1	p_{11}	p_{12}	P
x_2	p_{21}	p_{22}	Q
Total	P'	Q'	1

(*b*) Consider the following probability distribution :

$X \downarrow$ \ $Y \rightarrow$	0	1	2
0	0·1	0·2	0·1
1	0·2	0·3	0·1

Calculate $E(X)$, Var (X), Cov(X, Y) and $r\ (X, Y)$.

[***Delhi Univ. M.A. (Eco.), 1991***]

(*c*) Let (X, Y) have the p.m.f.

$$p(0, 1) = p(1, 0) = \frac{1}{3}\ ;\ p(0, -1) = p(-1, 0) = \frac{1}{6}.$$

Find $r(X, Y)$. Are X and Y independent ? For what values of $k, X + kY$ and $kX + Y$ are uncorrelated ?

10·5. Probable Error of Correlation Coefficient. If r is the correlation coefficient in a sample of n pairs of observations, then its *standard error* is given by

$$\text{S.E}(r) = \frac{1 - r^2}{\sqrt{n}}$$

Probable error of correlation coefficient is given by

$$\text{P.E.}(r) = 0{\cdot}6745 \times \text{S.E.} = 0{\cdot}6745 \frac{(1 - r^2)}{\sqrt{n}} \qquad \ldots(10{\cdot}6)$$

Probable error is an old measure for testing the reliability of an observed correlation coefficient. The reason for taking the factor 0·6745 is that in a normal distribution, the range $\mu \pm 0{\cdot}6745\ \sigma$ covers 50% of the total area. According to Secrist, "*The probable error of the correlation co-efficient is an amount which if added to and substracted from the mean correlation coefficient, produces amounts within which the chances are even that a coefficient of correlation from a series selected at random will fall.*"

If $r <$ P.E.(r), correlation is not at all significant. If $r > 6$ P.E.(r), it is definitely significant. A rigorous method (t-test) of testing the significance of an observed correlation coefficient will be discussed later in "*tests of significance*" in sampling [c.f. § 14·4·11].

Probable error also enables us to find the limits within which the population correlation coefficient can be expected to vary. The limits are $r \pm \text{P.E.}(r)$.

10·6. Rank Correlation. Let us suppose that a group of n individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. These ranks in the two characteristics will, in general, be different. For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also. Let (x_i, y_i); $i = 1, 2, \ldots, n$ be the ranks of the ith individual in two characteristics A and B respectively. Pearsonian coefficient of correlation between the ranks x_i's and y_i's is called the rank correlation coefficient between A and B for that group of individuals.

Assuming that no two individuals are bracketed equal in either classification, each of the variables X and Y takes the values $1, 2, \ldots, n$.

Hence $$\bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + \ldots + n) = \frac{n+1}{2}$$

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}(1^2 + 2^2 + \ldots + n^2) - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

$$\therefore \qquad \sigma_X^2 = \frac{n^2-1}{12} = \sigma_Y^2$$

In general $x_i \neq y_i$. Let $d_i = x_i - y_i$

$$\therefore \qquad d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad (\because \bar{x} = \bar{y})$$

Squaring and summing over i from 1 to n, we get

$$\sum d_i^2 = \sum\{(x_i - \bar{x}) - (y_i - \bar{y})\}^2$$

$$= \sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2 - 2\sum(x_i - \bar{x})(y_i - \bar{y})$$

Dividing both sides by n, we get

$$\frac{1}{n}\sum d_i^2 = \sigma_X^2 + \sigma_Y^2 - 2\,\text{Cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho\,\sigma_X\sigma_Y,$$

where ρ is the rank correlation coefficient between A and B.

$$\therefore \qquad \frac{1}{n}\sum d_i^2 = 2\sigma_X^2 - 2\rho\sigma_X^2 \quad \Rightarrow \quad 1 - \rho = \frac{\sum d_i^2}{2n\sigma_X^2}$$

$$\Rightarrow \qquad \rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_X^2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \ldots(10\cdot7)$$

which is the *Spearman's formula for the rank correlation coefficient.*

Remark. We always have

$$\sum d_i = \sum(x_i - y_i) = \sum x_i - \sum y_i = n(\bar{x} - \bar{y}) = 0 \qquad (\because \bar{x} = \bar{y})$$

This serves as a check on the calculations.

10·6·1. Tied Ranks. If some of the individuals receive the same rank in a ranking of merit, they are said to be tied. Let us suppose that m of the individuals, say, $(k+1)^{th}$, $(k+2)^{th}$, ..., $(k+m)^{th}$ are tied. Then each of these m individuals is assigned a common rank, which is the arithmetic mean of the ranks $k+1, k+2, \ldots, k+m$.

Derivation of $\rho(X, Y)$: We have :

$$\rho(X, Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{[\Sigma(X-\bar{X})^2 \,.\, \Sigma(Y-\bar{Y})^2]^{1/2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \,.\, \Sigma y^2}} \quad \ldots(^*)$$

where $\quad x = X - \bar{X},\ y = Y - \bar{Y}$.

If X and Y each takes the values 1, 2, ..., n, then we have

$$\bar{X} = (n+1)/2 = \bar{Y}$$

and $\quad n\sigma_X^2 = \Sigma x^2 = \dfrac{n(n^2-1)}{12}$ and $n\sigma_Y^2 = \Sigma y^2 = \dfrac{n(n^2-1)}{12} \quad \ldots(^{**})$

Also $\quad \Sigma d^2 = \Sigma(X-Y)^2 = \Sigma[(X-\bar{X})-(Y-\bar{Y})]^2 = \Sigma(x-y)^2$

$\Rightarrow \quad \Sigma d^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy$

$\Rightarrow \quad \Sigma xy = \dfrac{1}{2}[\Sigma x^2 + \Sigma y^2 - \Sigma d^2] \quad \ldots(^{***})$

We shall now investigate the effect of common ranking, (in case of ties), on the sum of squares of the ranks. Let S^2 and S_1^2 denote the sum of the squares of untied and tied ranks respectively.

Then we have :

$$\begin{aligned} S^2 &= (k+1)^2 + (k+2)^2 + \ldots + (k+m)^2 \\ &= mk^2 + (1^2 + 2^2 + \ldots + m^2) + 2k.(1 + 2 + \ldots + m) \\ &= mk^2 + \frac{m(m+1)(2m+1)}{6} + mk(m+1) \\ S_1^2 &= m\,(\text{Average rank})^2 \\ &= m\left[\frac{(k+1)+(k+2)+\ldots+(k+m)}{m}\right]^2 \\ &= m\left(k + \frac{m+1}{2}\right)^2 = mk^2 + \frac{m(m+1)^2}{4} + mk(m+1) \end{aligned}$$

$$\therefore \quad S^2 - S_1^2 = \frac{m(m+1)}{12}[2(2m+1) - 3(m+1)] = \frac{m(m^2-1)}{12}$$

Thus the effect of tying m individuals (ranks) is to reduce the sum of the squares by $m(m^2-1)/12$, though the mean value of the ranks remains the same, *viz.*, $(n+1)/2$.

Suppose that there are s such sets of ranks to be tied in the X-series so that the total sum of squares due to them is

$$\frac{1}{12}\sum_{i=1}^{s} m_i(m_i^2 - 1) = \frac{1}{12}\sum_{i=1}^{s}(m_i^3 - m_i) = T_X, \text{ (say)} \quad \ldots (10\cdot7a)$$

Similarly suppose that there are t such sets of ranks to be tied with respect to the other series Y so that sum of squares due to them is :

$$\frac{1}{12}\sum_{j=1}^{t} m_j' \cdot (m_j'^2 - 1) = \frac{1}{12}\sum_{j=1}^{t} (m_j'^3 - m_j') = T_Y \text{, (say)} \qquad \text{...(10·7b)}$$

Thus, in the case of ties, the new sums of squares are given by :

$$n\ \text{Var}'(X) = \sum x^2 - T_X = \frac{n(n^2-1)}{12} - T_X$$

$$n\ \text{Var}'(Y) = \sum y^2 - T_Y = \frac{n(n^2-1)}{12} - T_Y$$

and $$n\ \text{Cov}'(X, Y) = \frac{1}{2}[\sum x^2 - T_X + \sum y^2 - T_Y - \sum d^2] \qquad \text{[From (***)]}$$

$$= \frac{1}{2}\left[\frac{n(n^2-1)}{12} - T_X + \frac{n(n^2-1)}{12} - T_Y - \sum d^2\right]$$

$$= \frac{n(n^2-1)}{12} - \frac{1}{2}\left[(T_X + T_Y) + \sum d^2\right]$$

$$\rho(X, Y) = \frac{\dfrac{n(n^2-1)}{12} - \dfrac{1}{2}\left[T_X + T_Y + \sum d^2\right]}{\left[\dfrac{n(n^2-1)}{12} - T_X\right]^{1/2}\left[\dfrac{n(n^2-1)}{12} - T_Y\right]^{1/2}}$$

$$= \frac{\dfrac{n(n^2-1)}{6} - \left[\sum d^2 + T_X + T_Y\right]}{\left[\dfrac{n(n^2-1)}{6} - 2T_X\right]^{1/2}\left[\dfrac{n(n^2-1)}{6} - 2\,T_Y\right]^{1/2}} \qquad \text{...(10·7c)}$$

where T_X and T_Y are given by (10·7a) and (10·7b).

Remark. If we adjust only the covariance term *i.e.*, $\sum xy$ and not the variances σ_X^2 (or $\sum x^2$) and σ_Y^2 (or $\sum y^2$) for ties, then the formula (10·7c) reduces to :

$$\rho(X, Y) = \frac{\dfrac{n(n^2-1)}{6} - (\sum d^2 + T_X + T_Y)}{n(n^2-1)/6}$$

$$= 1 - \frac{6\,[\sum d^2 + T_X + T_Y]}{n(n^2-1)}, \qquad \text{...(10·7d)}$$

a formula which is commonly used in practice for numerical problems. For illustration, see Example 10·18.

Example 10·16. *The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics.*

(1, 1) (2, 10) (3, 3) (4, 4) (5, 5) (6, 7) (7, 2) (8, 6) (9, 8) (10, 11) (11, 15) (12, 9) (13, 14) (14, 12) (15, 16) (16, 13).

Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics.

Solution.

Ranks in Maths. (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics (Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
$d = X - Y$	0	−8	0	0	0	−1	5	2	1	−1	−4	3	−1	2	−1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Rank correlation coefficient is given by

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 136}{16 \times 255} = 1 - \frac{1}{5} = \frac{4}{5} = 0·8$$

Example 10·17. *Ten competitors in a musical test were ranked by the three judges A, B and C in the following order :*

Ranks by A :	1	6	5	10	3	2	4	9	7	8
Ranks by B :	3	5	8	4	7	10	2	1	6	9
Ranks by C :	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

Solution. Here $n = 10$

Ranks by A (X)	Ranks by B (Y)	Ranks by C (Z)	$d_1 = X-Y$	$d_2 = X-Z$	$d_3 = Y-Z$	d_1^2	d_2^2	d_3^2
1	3	6	− 2	− 5	−3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	− 3	− 4	−1	9	16	1
10	4	8	6	2	−4	36	4	16
3	7	1	− 4	2	6	16	4	36
2	10	2	− 8	0	8	64	0	64
4	2	3	2	1	−1	4	1	1
9	1	10	8	−1	−9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	− 1	1	2	1	1	4
Total			$\sum d_1 = 0$	$\sum d_2 = 0$	$\sum d_3 = 0$	$\sum d_1^2 = 200$	$\sum d_2^2 = 60$	$\sum d_3^2 = 214$

$$\rho(X, Y) = 1 - \frac{6 \sum d_1^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$\rho(X, Z) = 1 - \frac{6 \sum d_2^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho(Y, Z) = 1 - \frac{6\sum d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -\frac{49}{165}$$

Since $\rho(X, Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

10·6·2. Repeated Ranks (Continued). If any two or more individuals are bracketed equal in any classification with respect to characteristics A and B, or if there is more than one item with the same value in the series, then the Spearman's formula (10·7) for calculating the rank correlation coefficient breaks down, since in this case each of the variables X and Y does not assume the values $1, 2, \ldots, n$ and consequently, $\bar{x} \neq \bar{y}$.

In this case, common ranks are given to the repeated items. This common rank is the average of the ranks which these items would have assumed if they were sightly different from each other and the next item will get the rank next to the ranks already assumed. As a result of this, following adjustment or correction is made in the rank correlation formula [c.f. (10·7c) and (10·7d)].

In the formula, we add the factor $\dfrac{m(m^2-1)}{12}$ to $\sum d^2$, where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the X-series and Y-series.

Example 10·18. *Obtain the rank correlation coefficient for the following data:*

X	:	68	64	75	50	64	80	75	40	55	64
Y	:	62	58	68	45	81	60	68	48	50	70

Solution.

CALCULATIONS FOR RANK CORRELATION

X	Y	Rank X (x)	Rank Y (y)	$d = x - y$	d^2
68	62	4	5	−1	1
64	58	6	7	−1	1
75	68	2·5	3·5	−1	1
50	45	9	10	−1	1
64	81	6	1	5	25
80	60	1	6	−5	25
75	68	2·5	3·5	−1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\sum d = 0$	$\sum d^2 = 72$

In the X-series we see that the value 75 occurs 2 times. The common rank given to these values is 2·5 which is the average of 2 and 3, the ranks which these values would have taken if they were different. The next value 68, then gets the next rank which is 4. Again we see that value 64 occurs thrice. The common rank given to it is 6 which is the average of 5, 6 and 7. Similarly in the Y-series, the value 68 occurs twice and its common rank is 3·5 which is the average of 3 and 4. As a result of these common rankings, the formula for 'ρ' has to be corrected. To Σd^2 we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs. In the X-series the correction is to be applied twice, once for the value 75 which occurs twice ($m = 2$) and then for the value 64 which occurs thrice ($m = 3$). The total correction for the X-series is

$$\frac{2(4-1)}{12}+\frac{3(9-1)}{12}=\frac{5}{2}$$

Similarly, this correction for the Y-series is $\frac{2(4-1)}{12}=\frac{1}{2}$, as the value 68 occurs twice.

Thus $$\rho = 1-\frac{6\left[\Sigma d^2+\frac{5}{2}+\frac{1}{2}\right]}{n(n^2-1)} = 1-\frac{6(72+3)}{10\times 99}=0{\cdot}545$$

10·6·3. Limits for the Rank Correlation Coefficient. Spearman's rank correlation coefficient is given by

$$\rho = 1-\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

'ρ' is maximum, if $\sum_{i=1}^{n} d_i^2$ is minimum, *i.e.*, if each of the deviations d_i is minimum. But the minimum value of d_i is zero in the particular case $x_i = y_i$, *i.e.*, if the ranks of the ith individual in the two characteristic are equal. Hence the maximum value of ρ is $+1$, *i.e.*, $\rho \leq 1$.

'ρ' is minimum , if $\sum_{i=1}^{n} d_i^2$ is maximum, *i.e.*, if each of the deviations d_i is maximum which is so if the ranks of the n individuals in the two characteristics are in the opposite directions as given below :

x	1	2	3	...	...	$n-1$	n
y	n	$n-1$	$n-2$	...	...	2	1

...(*)

Case 1. Suppose n is odd and equal to $(2m+1)$ then the values of d are :

$d : 2m, 2m-2, 2m-4, \ldots, 2, 0, -2, -4, \ldots -(2m-2), -2m.$

$$\therefore \sum_{i=1}^{n} d_i^2 = 2\{(2m)^2+(2m-2)^2+\ldots+4^2+2^2\}$$

$$= 8\{m^2+(m-1)^2+\ldots+1^2\} = \frac{8m(m+1)(2m+1)}{6}$$

$$\text{Hence} \quad \rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{8m(m+1)(2m+1)}{(2m+1)\{(2m+1)^2-1\}}$$

$$= \frac{8m(m+1)}{(4m^2+4m)} = 1 - \frac{8m(m+1)}{4m(m+1)} = -1$$

Case II. Let n be even and equal to $2m$, (say).

Then the values of d are

$(2m-1), (2m-3), \ldots, 1, -1, -3, \ldots, -(2m-3), -(2m-1)$

$$\therefore \quad \Sigma d_i^2 = 2\{(2m-1)^2 + (2m-3)^2 + \ldots + 1^2\}$$

$$= 2[\{(2m)^2 + (2m-1)^2 + (2m-2)^2 + \ldots + 2^2 + 1^2\} - \{(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2\}]$$

$$= 2[1^2 + 2^2 + \ldots + (2m)^2 - \{2^2m^2 + 2^2(m-1)^2 + \ldots + 2^2\}]$$

$$= 2\left[\frac{2m(2m+1)(4m+1)}{6} - 4\,\frac{m(m+1)(2m+1)}{6}\right]$$

$$= \frac{2m}{3}[(2m+1)(4m+1) - 2(m+1)(2m+1)]$$

$$= \frac{2m}{3}[(2m+1)(4m+1-2m-2)]$$

$$= \frac{2m}{3}(2m+1)(2m-1) = \frac{2m(4m^2-1)}{3}$$

$$\therefore \quad \rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{4m(4m^2-1)}{2m(4m^2-1)} = -1$$

Thus the limits for rank correlation coefficient are given by $-1 \le \rho \le 1$.

Aliter. For an alternate and simpler proof for obtaining the minimum value of ρ, from (*) onward, proceed as in Hint to Question Number 9 of Exercise 10(c).

Remarks on Spearman's Rank Correlation Coefficient.

1. $\Sigma d = \Sigma x - \Sigma y = n(\bar{x} - \bar{y}) = 0$, which provides a check for numerical calculations.

2. Since Spearman's rank correlation coefficient ρ is nothing but Pearsonian correlation coefficient between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.

3. Karl Pearson's correlation coefficient assume that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure which is distribution free (or non-parametric). A distribution-free measure is one which does not make any assumptions about the parameters of the population. Spearman's ρ is such a measure (*i.e.*, distribution-free), since no strict assumptions are made about the form of the population from which sample observations are drawn.

4. Spearman' formula is easy to understand and apply as compared with Karl Pearson's formula. The value obtained by the two formulae, *viz.*, Pearsonian r and Spearman's ρ, are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is

always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.

6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution (Correlation Table). For $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time-consuming.

EXERCISE 10(c)

1. Prove that Spearman's rank correlation coefficient is given by

$1 - \dfrac{6 \sum d_i^2}{n^3 - n}$, where d_i denotes the difference between the ranks of ith individual.

2. (*a*) Explain the difference between product moment correlation coefficient and rank correlation coefficient.

(*b*) The rankings of ten students in two subjects A and B are as follows :

A :	3	5	8	4	7	10	2	1	6	9
B :	6	4	9	8	1	2	3	10	5	7

Find the correlation coefficient.

3. (*a*) Calculate the coefficient of correlation for ranks from the following data :

(X, Y) : (5, 8), (10, 3), (6, 2), (3, 9), (19, 12), (5, 3), (6, 17), (12, 18), (8, 22), (2, 12), (10, 17), (19, 20).

[*Calicut Univ. B.Sc. (Subs. Stat.), Oct. 1991*]

(*b*) Ten recruits were subjected to a selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test.

The marks secured by recruits in the selection test (X) and in the proficiency test (Y) are given below :—

Serial No. :	1	2	3	4	5	6	7	8	9	10
X :	10	15	12	17	13	16	24	14	22	20
Y :	30	42	45	46	33	34	40	35	39	38

Calculate product moment correlation coefficient and rank correlation coefficient. Why are two coefficients different ?

4. (*a*) The I.Q.'s of a group of 6 persons were measured, and they then sat for a certain examination. Their I.Q.'s and examination marks were as follows :

Person :	A	B	C	D	E	F
I.Q. :	110	100	140	120	80	90
Exam. marks :	70	60	80	60	10	20

Compute the coefficients of correlation and rank correlation. Why are the correlation figures obtained different ?

Ans. 0·882 and 0·9.

The difference arises due to the fact that when ranking is used instead of the full set of observations, there is always some loss of information.

(*b*) The value of ordinary correlation (r) for the following data is 0·636 :—

X :	·05	·14	·24	·30	·47	·52	·57	·61	·67	·72
Y :	1·08	1·15	1·27	1·33	1·41	1·46	1·54	2·72	4·01	9·63

(*i*) Calculate Spearman's rank-correlation (ρ) for this data.

(*ii*) What advantage of ρ was brought out in this example ?

4. Ten competitors in a beauty contest are ranked by three judges as follows :

		Competitors								
Judges :	1	2	3	4	5	6	7	8	9	10
A :	6	5	3	10	2	4	9	7	8	1
B :	5	8	4	7	10	2	1	6	9	3
C :	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges has the nearest approach to common tastes of beauty.

5. A sample of 12 fathers and their eldest sons gave the following data about their height in inches :

Father :	65	63	67	64	68	62	70	66	68	67	69	71
Son :	68	66	68	65	69	66	68	65	71	67	68	70

Calculate coefficient of rank correlation. (**Ans.** 0·7220)

6. The coefficient of rank correlation between marks in Statistics and marks in Mathematics obtained by a certain group of students is 0·8. If the sum of the squares of the difference in ranks is given to be 33, find the number of student in the group (**Ans.** 10). [*Madras Univ. B.Sc., 1990*]

7. The coefficient of rank correlation of the marks obtained by 10 students in Maths and Statistics was found to be 0·5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Hint. $$0{\cdot}5 = 1 - \frac{6\Sigma d^2}{10 \times 99}$$

$$\Rightarrow \qquad \Sigma d^2 = \frac{990}{6 \times 2} = 82{\cdot}5$$

Since one difference was wrongly taken as 3 instead of 7, the correct value of Σd^2 is given by

$$\text{Corrected } \Sigma d^2 = 82{\cdot}5 - (3)^2 + (7)^2 = 122{\cdot}5$$

$$\therefore \qquad \text{Corrected } \rho = 1 - \frac{6 \times 122{\cdot}5}{10 \times 99} = 0{\cdot}2576$$

8. If d_i be the difference in the ranks of the ith individual in two different characteristics, then show that the maximum value of $\sum_{i=1}^{n} d_i^2$ is $\frac{1}{3}(n^3 - n)$. Hence or otherwise, show that rank correlation coefficient lies between -1 and $+1$. [*Delhi Univ. B.Sc. (Maths. Hons.), 1986*]

9. Let $x_1, x_2, \ldots, x_n$ be the ranks of n individuals according to a character A and $y_1, y_2, \ldots, y_n$ be the ranks of the same individuals according to other character B. Obviously $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are permutations of 1, 2, ..., n. It is given that $x_i + y_i = 1 + n$, for $i = 1, 2, \ldots, n$. Show that the value of the rank correlation coefficient ρ between the characters A and B is -1.

Hint. We are given $\quad x_i + y_i = n + 1 \;\forall i = 1, 2, \ldots, n$

Also $\quad x_i - y_i = d_i$

$$\therefore \quad 2x_i = n + 1 + d_i \Rightarrow d_i = 2x_i - (n + 1)$$

$$\therefore \quad \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [4x_i^2 + (n+1)^2 - 2(n+1)2x_i]$$

$$= 4\,\frac{n(n+1)(2n+1)}{6} + n(n+1)^2 - \frac{4(n+1)n(n+1)}{2}$$

$$= \frac{n(n^2 - 1)}{3}$$

$$\therefore \quad \rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} = -1$$

Remark. From Spearmans' formula we note that ρ will be minimum if $\sum d_i^2$ is maximum, which will be so if the ranks X and Y are in opposite directions as given below :

x	1	2	3	...	n
y	n	$n-1$	$n-2$	...	1

This gives us

$$x_i + y_i = n + 1, \; i = 1, 2, \ldots, n.$$

Hence the value of $\rho = -1$ obtained above is minimum value of ρ.

10. Show that in a ranked bivariate distribution in which no ties occur and in which the variables are independent

(*a*) $\sum_i d_i^2$ is always even, and

(*b*) there are not more than $\frac{1}{6}(n^3 - n) + 1$ possible values of r.

11. Show that if X, Y be identically distributed with common probability mass function : $P(X = k) = \frac{1}{N}$, for $k = 1, 2, \ldots, N$; $N > 1$,

then $\rho_{X,Y}$, the correlation coefficient between X and Y, is given by

$$1 - \frac{6E(X-Y)^2}{N^2 - 1}$$

[*Delhi Univ. B.Sc. (Maths Hons.), 1992*]

10·7. Regression. The term "*regression*" literally means "*stepping back towards the average*". It was first used by a British biometrician Sir Francis Galton (1822—1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to "regress" or "step back" to the average population height. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry.

Definition. *Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable* and the variable which influences the values or is used for prediction, is called *independent variable.* In regression analysis independent variable is also known as *regressor or predictor* or *explanatory variable* while the dependent variable is also known as *regressed* or *explained variable.*

10·7·1. Lines of Regression. If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round some curve called the "*curve of regression*". If the curve is a straight line, it is called the line of regression and there is said to be *linear regression* between the variables, otherwise regression is said to be *curvilinear.*

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of "*best fit*" and is obtained by the *principles of least squares.*

Let us suppose that in the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$; Y is dependent variable and X is independent variable. Let the line of regression of Y on X be $Y = a + bX$.

According to the principle of least squares, the normal equations for estimating a and b are {*c.f.* (9·2*a*)}

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \qquad \ldots(10\cdot8)$$

and
$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \qquad \ldots(10\cdot9)$$

From (10·8) on dividing by n, we get

$$\bar{y} = a + b\bar{x} \qquad \ldots(10\cdot10)$$

Thus the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$.

Now

$$\mu_{11} = \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} \Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\,\bar{y} \qquad \ldots(10\cdot11)$$

Also
$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \quad \Rightarrow \quad \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \sigma_X^2 + \bar{x}^2 \qquad \ldots(10\cdot11a)$$

Dividing (10·9) by n and using (10·11) and (10·11a), we get

$$\mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2) \qquad \ldots(10\cdot12)$$

Multiplying (10·10) by $\bar{x}$ and then subtracting from (10·12), we get

$$\mu_{11} = b\sigma_X^2 \quad \Rightarrow \quad b = \frac{\mu_{11}}{\sigma_X^2} \qquad \ldots(10\cdot13)$$

Since 'b' is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x}) \qquad \ldots(10\cdot14)$$

$$\Rightarrow \qquad Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \qquad \ldots(10\cdot14a)$$

Starting with the equation $X = A + BY$ and proceeding similarly or by simply interchanging the variables X and Y in (10·14) and (10·14a), the equation of the line of regression of X on Y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y}) \qquad \ldots(10\cdot15)$$

$$\Rightarrow \qquad X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \qquad \ldots(10\cdot15a)$$

Aliter. The straight line defined by

$$Y = a + bX \qquad \ldots(i)$$

and satisfying the residual (least square) condition

$$S = E[(Y - a - bX)^2] = \text{Minimum} \qquad \ldots(ii)$$

for variations in a and b, is called the line of regression of Y on X.

The necessary and sufficient conditions for a minima of S, subject to variations in a and b are :

(i) $\dfrac{\partial S}{\partial a} = 0$, $\dfrac{\partial S}{\partial b} = 0$ and ...(*)

(ii) $\Delta = \begin{vmatrix} \dfrac{\partial^2 S}{\partial a^2} & \dfrac{\partial^2 S}{\partial a \partial b} \\ \dfrac{\partial^2 S}{\partial b\, \partial a} & \dfrac{\partial^2 S}{\partial b^2} \end{vmatrix} > 0$ and $\dfrac{\partial^2 S}{\partial a^2} > 0$ ... (**)

Using (*), we get

$$\frac{\partial S}{\partial a} = -2\, E\,[Y - a - bX] = 0 \qquad ...(iii)$$

$$\frac{\partial S}{\partial b} = -2\, E\,[X(Y - a - bX)] = 0 \qquad ...(iv)$$

$$\Rightarrow \quad E(Y) = a + bE(X) \; ...(v) \quad \text{and} \quad E(XY) = aE(X) + bE\,(X^2) \qquad ...(vi)$$

Equation (v) implies that the line (i) of regression of Y on X passes through the mean value $[E(X),\ E(Y)]$.

Multiplying (v) by $E(X)$ and substracting from (vi), we get

$$E(XY) - E(X)E(Y) = b[E(X^2) - \{E(X)\}^2]$$

$$\Rightarrow \quad \text{Cov}\,(X, Y) = b\,\sigma_X^2 \quad \Rightarrow \quad b = \frac{\text{Cov}\,(X, Y)}{\sigma_X^2} = \frac{r\sigma_Y}{\sigma_X} \qquad ...(vii)$$

Subtracting (v) from (i) and using (vii), we obtain the equation of line of regression of Y on X as :

$$Y - E(Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X^2}[X - E(X)] \;\Rightarrow\; Y - E(Y) = \frac{r\sigma_Y}{\sigma_X}[X - E(X)]$$

Similarly, the straight line defined by $X = A + BY$

and satisfying the residual condition

$$E[X - A - BY]^2 = \text{Minimum},$$

is called the line of regression of X on Y.

Remarks 1. We note that

$$\frac{\partial^2 S}{\partial a^2} = 2 > 0, \text{ and}$$

$$\frac{\partial^2 S}{\partial b^2} = 2E(X^2) \quad \text{and} \quad \frac{\partial^2 S}{\partial a \partial b} = 2E(X)$$

Substituting in (**), we have

$$\Delta = \frac{\partial^2 S}{\partial a^2} \cdot \frac{\partial^2 S}{\partial b^2} - \left(\frac{\partial^2 S}{\partial a \partial b} \right)^2$$

$$= 4\,[E(X^2) - (E(X))^2] = 4 .\, \sigma_X^2 > 0$$

Hence the solution of the least square equations (iii) and (iv), in fact, provides a minima of S.

2. The regression equation (10·14a) implies that the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$. Similarly (10·15a) implies that the line of regression of X on Y also passes through the point $(\bar{x}, \bar{y})$. Hence both the lines of regression pass through the point $(\bar{x}, \bar{y})$. In other words, the mean

values ($\bar{x}, \bar{y}$) can be obtained as the point of intersection of the two regression lines.

3. *Why two lines of Regression ?* There are always two lines of regression, one of Y on X and the other of X on Y. The line of regression of Y on X (10·14*a*) is used to estimate or predict the value of Y for any given value of X, *i.e.,* when Y is a dependent variable and X is an independent variable. The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of least squares. We can also obtain an estimate of X for any given value of Y by using equation (10·14*a*) but the estimate so obtained will not be best since (10·14*a*) is obtained on minimising the sum of the squares of errors of estimates in Y and not in X. Hence to estimate or predict X for any given value of Y, we use the regression equation of X on Y (10·15*a*) which is derived on minimising the sum of the squares of errors of estimates in X. Here X is a dependent variable and Y is an independent variable. The two regression equations are not reversible or interchangeable because of the simple reason that the basis and assumptions for deriving these equations are quite different. The regression equation of Y on X is obtained on minimising the sum of the squares of the errors parallel to the Y-axis while the regression equation of X on Y is obtained on minimising the sum of squares of the errors parallel to the X-axis.

In a particular case of perfect correlation, positive or negative, *i.e.*, $r \pm 1$, the equation of line of regression of Y on X becomes :

$$Y - \bar{y} = \pm \frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

$$\Rightarrow \quad \frac{Y - \bar{y}}{\sigma_Y} = \pm \left(\frac{X - \bar{x}}{\sigma_X}\right) \qquad \ldots(10{\cdot}16)$$

Similarly, the equation of the line of regression of X on Y becomes :

$$X - \bar{x} = \pm \frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

$$\Rightarrow \quad \frac{Y - \bar{y}}{\sigma_Y} = \pm \left(\frac{X - \bar{x}}{\sigma_X}\right),$$

which is same as (10·16).

Hence in case of perfect correlation, ($r = \pm 1$), both the lines of regression coincide. Therefore, in general, we always have two lines of regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

10·7·2. Regression Curves. In modern terminology, the conditional mean $E(Y \mid X = x)$ for a continuous distribution is called the regression function of Y on X and the graph of this function of x is known as the regression curve of Y on X or sometimes the regression curve for the mean of Y. Geometrically, the regression function represents the y co-ordinate of the centre of mass of the bivariate probability mass in the infinitesimal vertical strip bounded by x and $x + dx$.

Similarly, the regression function of X on Y is $E(X \mid Y = y)$ and the graph of this function of y is called the regression curve (of the mean) of X on Y.

In case a regression curve is a straight line, the corresponding regression is said to be *linear*. If one of the regressions is linear, it does not however follow that the other is also linear. For illustration, See Example 10·21.

Theorem 10·4. *Let (X, Y) be a two-dimensional random variable with $E(X) = \bar{X}$, $E(Y) = \bar{Y}$, $V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$ and let $r = r(X, Y)$ be the correlation coefficient between X and Y. If the regression of Y on X is linear then*

$$E(Y \mid X) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \qquad \ldots(10\text{·}16a)$$

Similarly, if the regression of X on Y is linear, then

$$E(X \mid Y) = \bar{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \qquad \ldots(10\text{·}16b)$$

Proof. Let the regression equation of Y on X be

$$E(Y \mid x) = a + bx \qquad \ldots(1)$$

But by definition,

$$E(Y \mid x) = \int_{-\infty}^{\infty} y f(y \mid x)\, dy = \int_{-\infty}^{\infty} y\, \frac{f(x, y)}{f_X(x)}\, dy$$

$$\therefore \qquad \frac{1}{f_X(x)} \int_{-\infty}^{\infty} y f(x, y)\, dy = a + bx \qquad \ldots(2)$$

Multiplying both sides of (2) by $f_X(x)$ and integrating w.r.t. x, we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x, y)\, dy dx = a\int_{-\infty}^{\infty} f_X(x)\, dx + b\int_{-\infty}^{\infty} x f_X(x) dx$$

$$\Rightarrow \qquad \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f(x, y)\, dx\right] dy = a + bE(X)$$

$$\Rightarrow \qquad \int_{-\infty}^{\infty} y f_Y(y) dy = a + bE(X)$$

i.e., $\qquad E(Y) = a + bE(X) \Rightarrow \bar{Y} = a + b\bar{X} \qquad \ldots(3)$

Multiplying both sides of (2) by $x f_X(x)$ and integrating w.r.t. x, we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f(x, y)\, dy\, dx = a\int_{-\infty}^{\infty} x f_X(x)\, dx + b\int_{-\infty}^{\infty} x^2 f_X(x)\, dx$$

$$\Rightarrow \qquad E(XY) = a\, E(X) + b\, E(X^2)$$

$$\Rightarrow \qquad \mu_{11} + \bar{X}\,\bar{Y} = a\bar{X} + b(\sigma_X^2 + \bar{X}^2) \qquad \ldots(4)$$

$[\because \mu_{11} = E(XY) - E(X)E(Y) = E(XY) - \overline{X}\,\overline{Y}$;

$$\sigma_X^2 = E(X^2) - \{E(X)\}^2 = E(X^2) - \overline{X}^2$$

Solving (3) and (4) simultaneously, we get

$$b = \frac{\mu_{11}}{\sigma_X^2} \quad \text{and} \quad a = \overline{Y} - \frac{\mu_{11}}{\sigma_X^2}\overline{X}$$

Substituting in (1) and simplifying, we get the required equation of the line of regression of Y on X as

$$E(Y \mid x) = \overline{Y} + \frac{\mu_{11}}{\sigma_X^2}(x - \overline{X})$$

$$\Rightarrow \quad E(Y \mid X) = \overline{Y} + \frac{\mu_{11}}{\sigma_X^2}(X - \overline{X})$$

$$\Rightarrow \quad E(Y \mid X) = \overline{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \overline{X})$$

By starting with the line $E(X \mid y) = A + By$ and proceeding similarly we shall obtain the equation of the line of regression of X on Y as

$$E(X \mid Y) = \overline{X} + \frac{\mu_{11}}{\sigma_Y^2}(Y - \overline{Y}) = \overline{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \overline{Y})$$

Example 10·19. *Given*

$$f(x, y) = xe^{-x(y+1)};\ x \geq 0,\ y \geq 0,$$

find the regression curve of Y on X. [B.H. Univ. M.Sc., 1989]

Solution. Marginal p.d.f. of X is given by

$$f_1(x) = \int_0^\infty f(x, y)\,dy = \int_0^\infty xe^{-x(y+1)}\,dy$$

$$= xe^{-x}\int_0^\infty e^{-xy}\,dy = xe^{-x}\left|\frac{e^{-xy}}{-x}\right|_0^\infty$$

$$= e^{-x},\ x \geq 0$$

Conditional p.d.f. of Y on X is given by

$$f(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{xe^{-x(y+1)}}{e^{-x}} = xe^{-xy},\ y \geq 0.$$

The regression curve of Y on X is given by

$$y = E(Y \mid X = x) = \int_0^\infty y\,f(y \mid x)\,dy = \int_0^\infty yxe^{-xy}\,dy$$

$$= x\left[\left| \frac{ye^{-xy}}{-x} \right|_0^\infty + \int_0^\infty \frac{e^{-xy}}{x} dy\right] = 0 + \left| \frac{e^{-xy}}{-x} \right|_0^\infty = \frac{1}{x}$$

i.e., $y = \frac{1}{x} \Rightarrow xy = 1.$

which is the equation of a rectangular hyperbola. Hence the regression of Y on X is not linear.

Example 10·20. *Obtain the regression equation of Y on X for the following distribution* :

$$f(x, y) = \frac{y}{(1 + x)^4} \exp\left(-\frac{y}{1 + x}\right) ; x, y \geq 0$$

Solution. Marginal p.d.f. of X is given by

$$f_1(x) = \int_0^\infty f(x, y)\, dy = \frac{1}{(1 + x)^4} \int_0^\infty y e^{-y/(1+x)} dy$$

$$= \frac{1}{(1 + x)^4} . \Gamma 2 . (1 + x)^2 \quad \text{(Using Gamma Integral)}$$

$$= \frac{1}{(1 + x)^2} ; \ x \geq 0$$

The conditional p.d.f. of Y (for given X) is

$$f(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{y}{(1 + x)^2} \exp\left(-\frac{y}{1 + x}\right) ; y \geq 0$$

Regression equation of Y on X is given by

$$y = E(Y \mid X) = \int_0^\infty y f(y \mid x)\, dy = \frac{1}{(1 + x)^2} \int_0^\infty y^2 e^{-y/(1+x)} dx$$

$$= \frac{1}{(1 + x)^2} . \ \Gamma 3 . \ (1 + x)^3 \quad \text{[Using Gamma Integral]}$$

$$\Rightarrow \quad y = 2(1 + x) \quad [\because \Gamma 3 = 2\,! = 2]$$

Hence the regression of Y on X is linear.

Example 10·21. *Let (X, Y) have the joint p.d.f. given by*

$$f(x, y) = \begin{cases} 1, \text{if } \mid y \mid < x, \ 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

Show that the regression of Y on X is linear but regression of X on Y is not linear.

Solution. $\mid y \mid < x \Rightarrow -x < y < x$ and $x > \mid y \mid$.

The marginal p.d.f's $f_1(.)$ of X and $f_2(.)$ of Y are given by :

$$f_1(x) = \int_{-x}^{x} f(x, y)\, dy = \int_{-x}^{x} 1 . dy = 2x ; 0 < x < 1$$

$$f_2(y) = \int_{\mid y \mid}^{1} f(x, y)\, dx = \int_{\mid y \mid}^{1} 1 . dx = 1 - \mid y \mid ; -1 < y < 1$$

$$\therefore \quad f_1(x \mid y) = \frac{f(x, y)}{f_2(y)} = \frac{1}{1 - |y|} \; ; -1 \le y < 1, 0 < x < 1$$

$$= \begin{cases} \dfrac{1}{1 - y}, & 0 < y < 1 \; ; \; 0 < x < 1 \\ \dfrac{1}{1 + y}, & -1 < y < 0 \; ; \; 0 < x < 1 \end{cases}$$

$$f_2(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{1}{2x}, \; 0 < x < 1; \; |y| < x$$

$$E(Y \mid X = x) = \int_{-x}^{x} y \cdot f_2(y \mid x)\, dy = \int_{-x}^{x} \frac{y}{2x}\, dy = \frac{1}{4x} \cdot \Big| y^2 \Big|_{-x}^{x} = 0$$

Hence the curve of regression of Y on X is $y = 0$, which is a straight line.

$$E(X \mid Y = y) = \int x f_1(x \mid y) dx$$

$$\therefore \quad E(X \mid Y = y) = \int_0^1 x \left(\frac{1}{1 - y} \right) dx = \frac{1}{2(1 - y)}, \; 0 < y < 1$$

and $$E(X \mid Y = y) = \int_0^1 x \left(\frac{1}{1 + y} \right) dx = \frac{1}{2(1 + y)}, \; -1 < y < 0$$

Hence the curve of regression of X on Y is

$$x = \begin{cases} \dfrac{1}{2(1 - y)}, & 0 < y < 1 \\ \dfrac{1}{2(1 + y)}, & -1 < y < 0, \end{cases}$$

which is not a straight line.

Example 10·22. *Variables X and Y have the joint p.d.f.*

$$f(x, y) = \frac{1}{3}(x + y), \; 0 \le x \le 1, \; 0 \le y \le 2.$$

Find :

(i) $r(X, Y)$

(ii) The two lines of regression

(iii) The two regression curves for the means.

Solution. The marginal p.d.f.'s of X and Y are given by :

$$f_1(x) = \int_0^2 f(x, y)\, dy = \frac{1}{3} \int_0^2 (x + y)\, dy = \frac{1}{3} \Big| xy + \frac{y^2}{2} \Big|_0^2$$

$$\Rightarrow \quad f_1(x) = \frac{2}{3}(1 + x) \; ; \; 0 \le x \le 1 \qquad \ldots(1)$$

$$f_2(y) = \int_0^1 f(x, y)\, dx = \frac{1}{3} \int_0^1 (x + y)\, dx = \frac{1}{3} \Big| \frac{x^2}{2} + xy \Big|_0^1$$

$$\Rightarrow \quad f_2(y) = \frac{1}{3}\left(\frac{1}{2} + y \right); \; 0 \le y \le 2 \qquad \ldots(2)$$

The conditional distributions are given by :

$$f_3(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{1}{2}\left(\frac{x+y}{1+x}\right)$$

$$f_4(x \mid y) = \frac{f(x, y)}{f_2(y)} = \frac{2(x+y)}{1+2y} \qquad \ldots(3)$$

$$E(Y|x) = \int_0^2 y \cdot f_3(y \mid x)\, dy = \frac{1}{2(1+x)} \int_0^2 y(x+y)dy$$

$$= \frac{1}{2(1+x)} \left| \frac{xy^2}{2} + \frac{y^3}{3} \right|_{y=0}^{y=2} = \frac{3x+4}{3(x+1)}$$

Similarly, we shall get

$$E(X \mid y) = \int_0^1 x f_4(x \mid y)\, dx = \frac{2}{1+2y} \int_0^1 (x^2 + xy)\, dx = \frac{2+3y}{3(1+2y)}$$

(*iii*) Hence the regression curves for means are :

$$y = E(Y|x) = \frac{3x+4}{3(x+1)} \quad \text{and} \quad x = E(X|y) = \frac{2+3y}{3(1+2y)}.$$

From the marginal distributions we shall get

$$E(X) = \int_0^1 x\, f_1(x)\, dx = \frac{5}{9}, \quad E(X^2) = \int_0^1 x^2 f_1(x) dx = \frac{7}{18}$$

$$\Rightarrow \quad \text{Var}(X) = \sigma_X^2 = \frac{7}{18} - \left(\frac{5}{9}\right)^2 = \frac{13}{162}$$

Similarly, we shall get

$$E(Y) = \frac{11}{9}, E(Y^2) = \frac{16}{9}; \sigma_Y^2 = \frac{16}{9} - \left(\frac{11}{9}\right)^2 = \frac{23}{81}$$

Also $\quad E(XY) = \int_0^1 \int_0^2 xy\, f(x, y)\, dxdy = \frac{1}{3} \int_0^1 \int_0^2 (x^2 y + xy^2)\, dxdy$

$$= \frac{1}{3}\left\{\left(\int_0^1 x^2 dx\right)\left(\int_0^2 y\, dy\right) + \left(\int_0^1 x\, dx\right)\left(\int_0^2 y^2 dy\right)\right\}$$

$$= \frac{1}{3}\left[\frac{1}{3} \times 2 + \frac{1}{2} \times \frac{8}{3}\right] = \frac{2}{3}$$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\, E(Y) = \frac{2}{3} - \frac{5}{9} \times \frac{11}{9} = -\frac{1}{81}$$

(*i*) $$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{-\frac{1}{81}}{\sqrt{\frac{13}{162} \times \frac{23}{81}}} = -\left(\frac{2}{299}\right)^{1/2}$$

(*ii*) The two lines of regression are :

$$Y - E(Y) = \frac{\text{Cov}(X, Y)}{\sigma_X^2}[X - E(X)] \quad \Rightarrow \quad Y - \frac{11}{9} = -\frac{2}{13}\left(X - \frac{5}{9}\right)$$

and $X - E(X) = \dfrac{\text{Cov}(X, Y)}{\sigma_Y^2}[Y - E(Y)] \quad \Rightarrow \quad X - \dfrac{5}{9} = -\dfrac{1}{23}\left(Y - \dfrac{11}{9}\right)$

10·7·3. Regression Coefficients. 'b', the slope of the line of regression of Y on X is also called the coefficient of regression of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. More precisely, we write

$$b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X} \qquad \text{...(10·17)}$$

Similarly, the coefficient of regression of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y and is given by

$$b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y} \qquad \text{...(10·17}a\text{)}$$

10·7·4. Properties of Regression Coefficients.

(*a*) *Correlation coefficient is the geometric mean between the regression coefficients.*

Proof. Multiplying (10·17) and (10·17*a*), we get

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2$$

$$\therefore \qquad r = \pm\sqrt{b_{XY} \times b_{YX}} \qquad \text{...(10·18)}$$

Remark. We have

$$r = \frac{\mu_{11}}{\sigma_X \cdot \sigma_Y}, \quad b_{YX} = \frac{\mu_{11}}{\sigma_X^2} \quad \text{and} \quad b_{XY} = \frac{\mu_{11}}{\sigma_Y^2}$$

It may be noted that the sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term μ_{11}. Thus if the regression coefficients are positive, 'r' is positive and if the regression coefficients are negative 'r' is negative.

From (10·18), we have

$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

the sign to be taken before the square root is that of the regression coefficients.

(*b*) *If one of the regression coefficients is greater than unity, the other must be less than unity.*

Proof. Let one of the regression coefficients (say) b_{YX} be greater than unity, then we have to show that $b_{XY} < 1$.

Now $\qquad b_{YX} > 1 \Rightarrow \dfrac{1}{b_{YX}} < 1 \qquad$ ...(*)

Also $\qquad r^2 \le 1 \quad \Rightarrow \quad b_{YX} \cdot b_{XY} \le 1$

Hence $\qquad b_{XY} \le \dfrac{1}{b_{YX}} < 1 \qquad$ [From (*)]

(*c*) *Arithmetic mean of the regression coefficients is greater than the correlation coefficient r, provided r > 0.*

Proof. We have to prove that $\frac{1}{2}(b_{YX} + b_{XY}) \geq r$

or $\quad \frac{1}{2}\left(r\,\frac{\sigma_Y}{\sigma_X} + r\,\frac{\sigma_X}{\sigma_Y}\right) \geq r \quad$ or $\quad \frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y} \geq 2 \qquad (\because\ r > 0)$

$\Rightarrow \quad \sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y \geq 0 \quad$ *i.e.,* $\quad (\sigma_Y - \sigma_X)^2 \geq 0$

which is always true, since the square of a real quantity is ≥ 0.

(*d*) *Regression coefficients are independent of the change of origin but not of scale.*

Proof. Let $U = \frac{X - a}{h}$, $V = \frac{Y - b}{k} \quad \Rightarrow \quad X = a + hU, Y = b + kV,$

where $a, b, h\ (> 0)$ and $k\ (> 0)$ are constants.

Then $\text{Cov}\,(X, Y) = hk\ \text{Cov}\,(U, V), \sigma_X^2 = h^2\sigma_U^2$ and $\sigma_Y^2 = k^2\sigma_V^2$

$$b_{YX} = \frac{\mu_{11}}{\sigma_X^2} = \frac{hk\ \text{cov}\,(U, V)}{h^2\sigma_U^2}$$

$$= \frac{k}{h} \cdot \frac{\text{cov}\,(U, V)}{\sigma_U^2} = \frac{k}{h} b_{VU}$$

Similarly, we can prove that

$$b_{XY} = (h/k)\, b_{UV}$$

10·7·5. Angle Between Two Lines of Regression. Equations of the lines of regression of Y on X, and X on Y are

$$Y - \bar{y} = r\,.\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \text{ and } X - \bar{x} = r\,.\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

Slopes of these lines are $r\,.\,\frac{\sigma_Y}{\sigma_X}$ and $\frac{\sigma_Y}{r\sigma_X}$ respectively. If θ is the angle between the two lines of regression then

$$\tan\theta = \frac{r\,\frac{\sigma_Y}{\sigma_X} \sim \frac{\sigma_Y}{r\sigma_X}}{1 + r\,\frac{\sigma_Y}{\sigma_X}\,.\,\frac{\sigma_Y}{r\sigma_X}} = \frac{r^2 \sim 1}{r} \cdot \left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

$$= \frac{1 - r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right) \qquad (\because\ r^2 \leq 1)$$

$$\therefore \quad \theta = \tan^{-1}\left\{\frac{1 - r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)\right\} \qquad \ldots(10\cdot19)$$

Case (i). ($r = 0$). If $r = 0$, $\tan\theta = \infty \quad \Rightarrow \quad \theta = \frac{\pi}{2}$

Thus if the two variables are uncorrelated, the lines of regression become perpendicular to each other.

Case (ii). ($r = \pm 1$). If $r = \pm 1$, $\tan\theta = 0 \quad \Rightarrow \quad \theta = 0$ or π.

In this case the two lines of regression either coincide or they are parallel to each other. But since both the lines of regression pass through the point

$(\bar{x}, \bar{y})$, they cannot be parallel. Hence in the case of perfect correlation, positive or negative, the two lines of regression coincide.

Remarks 1. Whenever two lines intersect, there are two angles between them, one acute-angle and the other obtuse angle. Further $\tan\theta > 0$ if $0 < \theta < \pi/2$, *i.e.*, θ is an acute angle and $\tan\theta < 0$ if $\pi/2 < \theta < \pi$, *i.e.*, θ is an obtuse angle and since $0 < r^2 < 1$, the acute angle (θ_1) and obtuse angle θ_2 between the two lines of regression are given by

$$\theta_1 = \text{Acute angle} = \tan^{-1}\left\{\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2} \cdot \frac{1 - r^2}{r}\right\}, r > 0$$

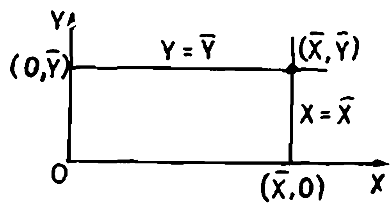
and

$$\theta_2 = \text{Obtuse angle} = \tan^{-1}\left\{\frac{\sigma_X \cdot \sigma_Y}{\sigma_X^2 + \sigma_Y^2} \cdot \frac{r^2 - 1}{r}\right\}, r > 0$$

2. When $r = 0$, *i.e.*, variables X and Y are uncorrelated, then the lines of regressions of Y on X and X on Y are given respectively by : [From (10·14a) and (10·15a)]

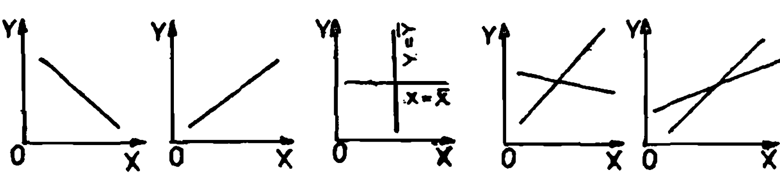
$$Y = \bar{Y} \text{ and } X = \bar{X},$$

as shown in the adjoining diagram. Hence, in this case ($r = 0$), the lines of regression are perpendicular to each other and are parallel to X- axis and Y-axis respectively.



3. The fact that if $r = 0$ (variables uncorrelated), the two lines of regression are perpendicular to each and if $r = \pm 1$, $\theta = 0$, *i.e.*, the two lines coincide, leads us to the conclusion that for higher degree of correlation between the variables, the angle between the lines is smaller, *i.e.*, the two lines of regression are nearer to each other. On the other hand, if the lines of regression make a larger angle, they indicate a poor degree of correlation between the variables and ultimately for $\theta = \pi/2$, $r = 0$, *i.e.*, the lines become perpendicular if no correlation exists between the variables. Thus by plotting the lines of regression on a graph paper, we can have an approximate idea about the degree of correlation between the two variables under study. Consider the following illustrations :

TWO LINES COINCIDE ($r = -1$)	TWO LINES COINCIDE ($r = +1$)	TWO LINES PERPENDICULAR ($r = 0$)	TWO LINES APART (LOW DEGREE OF CORRELATION)	TWO LINES APART (HIGH DEGREE OF CORRELATION



10·7·6. Standard Error of Estimate or Residual Variance. The equation of the line of regression of Y on X is

$$Y = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$\Rightarrow \qquad \frac{Y - \bar{Y}}{\sigma_Y} = r\left(\frac{X - \bar{X}}{\sigma_X}\right)$$

The residual variance s_Y^2 is the expected value of the squares of deviations of the observed values of Y from the expected values as given by the line of regression of Y on X. Thus

$$s_Y^2 = E[Y - \{\bar{Y} + (r\sigma_Y (X - \bar{X})/\sigma_X)\}]^2$$

$$= \sigma_Y^2 E\left[\frac{Y - \bar{Y}}{\sigma_Y} - r\left(\frac{X - \bar{X}}{\sigma_X}\right)\right]^2 = \sigma_Y^2 E(Y^* - rX^*)^2$$

where X^* and Y^* are standardised variates so that

$$E(X^{*2}) = 1 = E(Y^{*2}) \text{ and } E(X^* Y^*) = r.$$

$$\therefore \qquad s_Y^2 = \sigma_Y^2[E(Y^{*2}) + r^2 E(X^{*2}) - 2r E(X^*Y^*)] = \sigma_Y^2 (1 - r^2)$$

$$\Rightarrow \qquad s_Y = \sigma_Y (1 - r^2)^{1/2}$$

Similarly, the standard error of estimate of X is given by

$$s_X = \sigma_X (1 - r^2)^{1/2}$$

Remarks 1. Since s_X^2 or $s_Y^2 \geq 0$, it follows that

$$(1 - r^2) \geq 0 \Rightarrow |r| \leq 1$$

Hence
$$-1 \leq r(X, Y) \leq 1$$

2. If $r = \pm 1$, $s_X = s_Y = 0$ so that each deviation is zero, and the two lines of regression are coincident.

3. Since, as $r^2 \to 1$, s_X and $s_Y \to 0$, it follows that departure of tne value r^2 from unity indicates the departure of the relationship between the variables X and Y from linearity.

4. From the definition of linear regression, the minima condition implies that s_Y or s_X is the minimum variance.

10·7·7. Correlation Coefficient between Observed and Estimated Value. Here we will find the correlation between Y and

$$\hat{Y} = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

where $\hat{Y}$ is the estimated value of Y as given by the line of regression of Y on X, which is given by

$$r(Y, \hat{Y}) = \frac{\text{Cov}(\hat{Y}, Y)}{\sigma_Y \sigma_{\hat{Y}}}$$

We have

$$E(\hat{Y}) = E\left[\bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right] = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}E(X - \bar{X}) = \bar{Y}$$

$$\therefore \qquad \text{Var}(\hat{Y}) = E[\hat{Y} - E(\hat{Y})]^2 = E\left[r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right]^2 = r^2\sigma_Y^2$$

$$\Rightarrow \qquad \hat{\sigma}_Y = r\,\sigma_Y$$

Also $\text{Cov}(Y, \hat{Y}) = E[\{Y - E(Y)\}\ \{\hat{Y} - E(\hat{Y})\}]$

$$= E\left[\{b(X - E(X))\}\left\{r\,\frac{\sigma_Y}{\sigma_X}(X - \overline{X})\right\}\right]$$

$$= br\frac{\sigma_Y}{\sigma_X}E[\{X - E(X)\}^2] = \left(r\,\frac{\sigma_Y}{\sigma_X}\right)^2\sigma_X^2 = r^2\sigma_Y^2$$

$$\therefore \qquad r(Y, \hat{Y}) = \frac{r^2\sigma_Y^2}{\sigma_Y\, r\,\sigma_Y} = r = r(X, Y)$$

Hence the correlation coefficient between observed and estimated value of Y is the same as the correlation coefficient between X and Y.

Example 10·23. *Obtain the equations of the lines of regression for the data in Example 10·1. Also obtain the estimate of X for $Y = 70$.*

Solution. Let $U = X - 68$ and $V = Y - 69$, then

$$\bar{U} = 0,\ \bar{V} = 0,\ \sigma_U^2 = 4{\cdot}5,\ \sigma_V^2 = 5{\cdot}5,\ \text{Cov}(U, V) = 3 \text{ and } r(U, V) = 0{\cdot}6$$

Since correlation coefficient is independent of change of origin, we get

$$r = r(X, Y) = r(U, V) = 0{\cdot}6$$

We know that if $U = \dfrac{X - a}{h}$, $V = \dfrac{Y - b}{k}$, then

$$\bar{X} = a + h\bar{U},\ \bar{Y} = b + k\bar{V},\ \sigma_X = h\sigma_U \text{ and } \sigma_Y = k\,\sigma_Y$$

In our case $h = k = 1$, $a = 68$ and $b = 69$.

Thus $\quad \bar{X} = 68 + 0 = 68,\ \bar{Y} = 69 + 0 = 69$

$$\sigma_X = \sigma_U = \sqrt{4{\cdot}5} = 2{\cdot}12 \text{ and } \sigma_Y = \sigma_V = \sqrt{5{\cdot}5} = 2{\cdot}35$$

Equation of line of regression of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

i.e., $Y = 69 + 0{\cdot}6 \times \dfrac{2{\cdot}35}{2{\cdot}12}(X - 68) \Rightarrow Y = 0{\cdot}665\,X + 23{\cdot}78$

Equation of line of regression of X on Y is

$$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

$$\Rightarrow \qquad X = 68 + 0{\cdot}6 \times \frac{2{\cdot}12}{2{\cdot}35}(Y - 69) \quad \textit{i.e.,}\ X = 0{\cdot}54Y + 30{\cdot}74$$

To estimate X for given Y, we use the line of regression of X on Y. If $Y = 70$, estimated value of X is given by

$$\hat{X} = 0{\cdot}54 \times 70 + 30{\cdot}74 = 68{\cdot}54,$$

where $\hat{X}$ is estimate of X.

Example 10·24. *In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible :*

Variance of X = 9.

Regression equations : $8X - 10Y + 66 = 0,\ 40X - 18Y = 214.$

What were (i) *the mean values of X and Y,*

(ii) *the correlation coefficient between X and Y, and*

(iii) *the standard deviation of Y ?*

[Punjab Univ. B.Sc. (Hons.), 1993]

Solution (*i*) Since both the lines of regression pass through the point $(\overline{X}, \overline{Y})$, we have $8\overline{X} - 10\overline{Y} + 66 = 0$, and $40\overline{X} - 18\overline{Y} = 214$.

Solving, we get $\overline{X} = 13$, $\overline{Y} = 17$.

(*ii*) Let $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ be the lines of regression of Y on X and X on Y respectively. These equations can be put in the form :

$$Y = \frac{8}{10}X + \frac{66}{10} \text{ and } X = \frac{18}{40}Y + \frac{214}{40}$$

$\therefore \quad b_{YX}$ = Regression coefficient of Y on $X = \dfrac{8}{10} = \dfrac{4}{5}$

and $\quad b_{XY}$ = Regression coefficient of X on $Y = \dfrac{18}{40} = \dfrac{9}{20}$

Hence $\quad r^2 = b_{YX} \cdot b_{XY} = \dfrac{4}{5} \cdot \dfrac{9}{20} = \dfrac{9}{25}$

$\therefore \quad r = \pm \dfrac{3}{5} = \pm 0{\cdot}6$

But since both the regression coefficients are positive, we take $r = +0{\cdot}6$

(*iii*) We have $\quad b_{YX} = r \cdot \dfrac{\sigma_Y}{\sigma_X} \quad \Rightarrow \quad \dfrac{4}{5} = \dfrac{3}{5} \times \dfrac{\sigma_Y}{3} \quad [\because \sigma_X^2 = 9 \text{ (Given)}]$

Hence $\quad \sigma_Y = 4$

Remarks. 1. It can be verified that the values of $\overline{X} = 13$ and $\overline{Y} = 17$ as obtained in part (*i*) satisfy both the regression equations. In numerical problems of this type, this check should invariably be applied to ascertain the correctness of the answer.

2. If we had assumed that $8X - 10Y + 66 = 0$, is the equation of the line of regression of X on Y and $40X - 18Y = 214$ is the equation of line of regression of Y on X, then we get respectively :

$$8X = 10Y - 66 \text{ and } 18Y = 40X - 214$$

$$\Rightarrow \quad X = \frac{10}{8}Y - \frac{66}{8} \text{ and } Y = \frac{40}{18}X - \frac{214}{18}$$

$$\Rightarrow \quad b_{XY} = \frac{18}{8} \text{ and } b_{YX} = \frac{40}{18}$$

$$\therefore \quad r^2 = b_{XY} \cdot b_{YX} = \frac{10}{8} \times \frac{40}{18} = 2{\cdot}78$$

But since r^2 always lies between 0 and 1, our supposition is wrong.

Example 10·25. *Find the most likely price in Bombay corresponding to the price of Rs. 70 at Calcutta from the following :*

	Calcutta	*Bombay*
Average price	*65*	*67*
Standard deviation	*2·5*	*3·5*

Correlation coefficient between the prices of commodities in the two cities is 0·8. **[Nagpur Univ. B.Sc., 1993; Sri Venkateswara Univ. B.Sc. (Oct.) 1990]**

Solution. Let the prices, (in Rupees), in Bombay and Calcutta be denoted by Y and X respectively. Then we are given

$\bar{X} = 65, \bar{Y} = 67, \sigma_X = 2{\cdot}5, \sigma_Y = 3{\cdot}5$ and $r = r(X, Y) = 0{\cdot}8$. We want Y for $X = 70$.

Line of regression of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$\Rightarrow \qquad Y = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}(X - 65)$$

When $X = 70$, $\quad \hat{Y} = 67 + 0{\cdot}8 \times \dfrac{3{\cdot}5}{2{\cdot}5}(70 - 65) = 72{\cdot}6$

Example 10·26. *Can $Y = 5 + 2{\cdot}8\,X$ and $X = 3 - 0{\cdot}5Y$ be the estimated regression equations of Y on X and X on Y respectively ? Explain your answer with suitable theoretical arguments.* **[Delhi Univ. M.A.(Eco.), 1986]**

Solution. Line of regression of Y on X is :

$$Y = 5 + 2{\cdot}8X \quad \Rightarrow \quad b_{YX} = 2{\cdot}8 \qquad \ldots(*)$$

Line of regression of X on Y is :

$$X = 3 - 0{\cdot}5Y \quad \Rightarrow \quad b_{XY} = -\,0{\cdot}5 \qquad \ldots(**)$$

This is not possible, since each of the regression coefficients b_{YX} and b_{XY} must have the same sign, which is same as that of Cov (X, Y). If Cov (x, y) is positive, then both the regression coefficients are positive and if Cov (X, Y) is negative, then both the regression coefficients are negative. Hence (*) and (**) cannot be the estimated regression equations of Y on X and X on Y respectively.

EXERCISE 10 (d)

1. (*a*) Explain what are regression lines. Why are there two such lines ? Also derive their equations.

(*b*) Define (*i*) Line of regression, (*ii*) Regression coefficient. Find the equations to the lines of regression and show that the coefficient of correlation is the geometric mean of coefficients of regression.

(*c*) What equation is the equivalent mathematical statement for the following words ?

"If the respective deviations in each series, X and Y, from their means were expressed in units of standard deviations, *i.e.*, if each were divided by the standard deviation of the series, to which it belongs and plotted to a scale of standard deviations, the slope of a straight line best describing the plotted points would be the correlation coefficient r."

2(a) Obtain the equation of the line of regression of Y on X and show that the angle θ, between the two lines of regression is given by

$$\tan\,\theta = \frac{1-\rho^2}{\rho} \times \frac{\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}$$

where σ_1, σ_2 are the standard deviations of X and Y respectively, and ρ is the correlation coefficient. ***(Delhi Univ. B.Sc. (Maths. Hons.), 1989)***

Interpret the cases when $\rho = 0$ and $\rho = \pm 1$.

(Bangalore Univ. B.Sc. 1990)

(b) If θ is the acute angle between the two regression lines with correlation coefficient r, show that $\sin\theta \le 1 - r^2$.

3. (a) Explain the term "regression" by giving examples. Assuming that the regression of Y on X is linear, outline a method for the estimation of the coefficients in the regression line based on the random paired sample of X and Y, and show that the variance of the error of the estimate for Y for the regression line is $\sigma_Y^2\,(1-\rho^2)$, where σ_Y^2 is the variance of Y and ρ is the correlation coefficient between X and Y.

(b) Prove that X and Y are linearly related if and only if $\rho_{XY}^2 = 1$. Further show that the slope of the regression line is positive or negative according as $\rho = +1$ or $\rho = -1$.

(c) Let X and Y be two variates. Define $X^* = \dfrac{X-a}{b}$, $Y^* = \dfrac{Y-c}{d}$ for some constants a, b, c and d. Show that the regression line (least square) of Y on X can be obtained from that of Y^* on X^*.

(d) Show that the coefficient of correlation between the observed and the estimated values of Y obtained from the line of regression of Y on X, is the same as that between X and Y.

4. Two variables X and Y are known to be related to each other by the relation $Y = X/(aX + b)$. How is the theory of linear regression to be employed to estimate the constants a and b from a set of n pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$?

Hint. $$\frac{1}{Y} = \frac{aX+b}{X} = a + \frac{b}{X}$$

Put $$\frac{1}{X} = U \text{ and } \frac{1}{Y} = V$$

$\therefore$ $$V = a + bU$$

5. Derive the standard error of estimate of Y obtained from the linear regression equation of Y on X. What does this standard error measure ?

6. (a) Calculate the coefficient of correlation from the following data :

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

Also obtain the equations of the lines of regression and obtain an estimate of Y which should correspond on the average to $X = 6{\cdot}2$.

Ans. $r = 0{\cdot}95$, $Y - 12 = 0{\cdot}95\ (X - 5)$, $X - 5 = 0{\cdot}95\ (Y - 12)$, $13{\cdot}14$

(*b*) Why do we have, in general, two lines of regression ? Obtain the regression of Y on X, and X on Y from the following table and estimate the blood pressure when the age is 45 years :

Age in years (X)	*Blood pressure* (Y)	*Age in years* (X)	*Blood pressure* (Y)
56	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	152
47	128	60	155

Ans. $Y = 1{\cdot}138X + 80{\cdot}778$, $Y = 131{\cdot}988$ for $X = 45$.

(*c*) Suppose the observations on X and Y are given as :

X :	59	65	45	52	60	62	70	55	45	49
Y :	75	70	55	65	60	69	80	65	59	61

where $N = 10$ students, and Y = Marks in Maths, X = Marks in Economics. Compute the least square regression equations of Y on X and of X on Y.

If a student gets 61 marks in Economics, what would you estimate his marks in Maths to be ?

7. (*a*) In a correlation analysis on the ages of wives and husbands, the following data were obtained. Find

(*i*) the value of the correlation coefficient, and (*ii*) the lines of regression.

Estimate the age of husband whose wife's age is 31 years. Estimate the age of wife whose husband is 40 years old.

Age of wife → / Age of Husband ↓	15—25	25—35	35—45	45—55	55—65
15—30	30	6	3	—	—
30—45	18	32	15	12	8
45—60	2	28	40	16	9
60—75	—	4	9	10	8

(*b*) The following table gives the distribution of total cultivable area (X) and area under cultivation (Y) in a district of 69 villages.

Calculate (*i*) the linear regression of Y on X, (*ii*) the correlation coefficient $r(X, Y)$, and (*iii*) the average area under wheat corresponding to total area of 1,000 Bighas,

		Total area (in Bighas)				
		0—500	500—1000	1000—1500	1500—2000	2000—2500
Area under wheat	0—200	12	6	...	...	...
	200—400	2	18	4	2	1
	400—600	...	4	7	3	...
	600—800	...	1	...	2	1
	800—1000	...	...	1	2	3

Ans. (*i*) $Y = 0{\cdot}7641X - 455{\cdot}3854$, (*ii*) $r(X, Y) = 0{\cdot}756$
(*iii*) $Y = 308{\cdot}7146$ for $X = 1000$

8. (*a*) Compare and contrast the roles of correlation and regression in studying the inter-dependence of two variates.

For 10 observations on price (X) and supply (Y) the following data were obtained (in appropriate units).

$$\Sigma X = 130,\ \Sigma Y = 220,\ \Sigma X^2 = 2288,\ \Sigma Y^2 = 5506 \text{ and } \Sigma XY = 3467$$

Obtain the line of regression of Y on X and estimate the supply when the price is 16 units, and find out the standard error of the estimate.

Ans. $Y = 8{\cdot}8 + 1{\cdot}015X$, $25{\cdot}04$

(*b*) If a number X is chosen at random from among the integers 1, 2, 3, 4 and a number Y is chosen from among those at least as large as X, prove that

$$\text{Cov}\ (X, Y) = \frac{5}{8}$$

Find also the regression line of X on Y.

(*c*) Calculate the correlation coefficient from the following data :—

$$N = 100,\quad \Sigma X = 12500\quad \Sigma Y = 8000$$
$$\Sigma X^2 = 1585000,\quad \Sigma Y^2 = 648100\quad \Sigma XY = 1007425.$$

Also obtain the regression equation of Y on X.

9. (*a*) The means of a bivariate frequency distribution are at (3, 4) and $r = 0{\cdot}4$. The line of regression of Y on X is parallel to the line $Y = X$. Find the two lines of regression and estimate the mean of X when $Y = 1$.

(*b*) For certain data, $Y = 1{\cdot}2\ X$ and $X = 0{\cdot}6\ Y$, are the regression lines. Compute $\rho(X, Y)$ and σ_X/σ_Y. Also compute $\rho\ (X, Z)$, if $Z = Y - X$.

(*c*) The equations of two regression lines obtained in a correlation analysis are as follows :

$$3X + 12Y = 19,\quad 3Y + 9X = 46$$

Obtain (*i*) the value of correlation coefficient,
(*ii*) mean values of X and Y, and

(*iii*) the ratio of the coefficient of variability of X to that of Y.

Ans. (*i*) $-\frac{1}{2}\sqrt{3}$, (*ii*) $\bar{X} = 5, \bar{Y} = 1/3$.

(*d*) For an army personnel of strength 25, the regression of weight of kidneys (Y) on weight of heart (X), both measured in ounces is

$$Y - 0{\cdot}399X - 6{\cdot}934 = 0$$

and the regression of weight of heart on weight of kidney is

$$X - 1{\cdot}212Y + 2{\cdot}461 = 0$$

Find the correlation coefficient between X and Y and their mean values. Can you find out the standard deviation of X and Y as well ?

Ans. $r(X, Y) = 0{\cdot}70$, $\bar{X} = 11{\cdot}5086$, $\bar{Y} = 11{\cdot}5261$, No.

(*e*) Find the coefficient of correlation for distribution in which

$$\text{S.D of } X = 3{\cdot}0 \text{ units}$$
$$\text{S.D. of } Y = 1{\cdot}4 \text{ units}$$

Coefficient of regression of Y on $X = 0{\cdot}28$.

10. (*a*) Given that $X = 4Y + 5$ and $Y = kX + 4$, are the lines of regression of X on Y and Y on X respectively, show that $0 < 4k < 1$. If $k = \frac{1}{16}$, find the means of the two variables and coefficient of correlation between them.

[*Punjab Univ. B.Sc. (Hons.), 1989*]

Hint. $X = 4Y + 5 \Rightarrow b_{XY} = 4$

$Y = kx + 4 \Rightarrow b_{YX} = k$

$\therefore \quad r^2 = 4k \qquad \ldots(*)$

But $0 \leq r^2 \leq 1 \Rightarrow 0 \leq 4k \leq 1$.

If $k = \frac{1}{16}$, then from (*), we get

$r^2 = 4 \times \frac{1}{16} \Rightarrow r = +\frac{1}{2}$ [Since both the regression coefficient are positive]

For $k = \frac{1}{16}$, the two lines of regression become

$$X = 4Y + 5 \text{ and } Y = \frac{1}{16}X + 4$$

Solving the two equations, we get $\bar{Y} = 5{\cdot}75$, $\bar{X} = 28$.

(*b*) For 50 students of a class the regression equation of marks in Statistics (X) on marks in Mathematics (Y) is $3Y - 5X + 180 = 0$. The mean marks in Mathematics is 44 and variance of marks in Statistics is 9/16th of the variance of marks in Mathematics. Find the mean marks in Statistics and the coefficient of correlation between marks in two subjects.

[*Bangalore Univ. B.Sc., 1989*]

Hint. We are given $n = 50$, $\bar{Y} = 44$

and $\sigma_X^2 = \frac{9}{16}\sigma_Y^2 \Rightarrow \frac{\sigma_X}{\sigma_Y} = \frac{3}{4} \qquad \ldots(*)$

The equation of the line of regression of X on Y is given to be

$$3Y - 5X + 180 = 0 \quad \Rightarrow \quad X = \frac{3}{5}Y + \frac{180}{5}$$

$$\therefore \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{3}{5} \quad \Rightarrow \quad r \cdot \frac{3}{4} = \frac{3}{5} \quad \text{or} \quad r = 0{\cdot}8$$

Since the lines of regression pass through the point $(\bar{X}, \bar{Y})$, we get

$$\bar{X} = \frac{3}{5}\bar{Y} + \frac{180}{5} = \frac{3}{5} \times 44 + 36 = 62{\cdot}4$$

(*c*) Out of the two lines of regression given by

$$X + 2Y - 5 = 0 \text{ and } 2X + 3Y - 8 = 0,$$

which one is the regression line of X on Y?

Use the equations to find the mean of X and the mean of Y. If the variance of X is 12, calculate the variance of Y.

Ans. $\bar{X} = 1, \bar{Y} = 2, \sigma_Y^2 = 4$

(*d*) The lines of regression in a bivariate distribution are :

$$X + 9Y = 7 \text{ and } Y + 4X = \frac{49}{3}$$

Find (*i*) the coefficient of correlation, (*iii*) the ratios $\sigma_X^2 : \sigma_Y^2 : \text{Cov}(X, Y)$, (*iii*) the means of the distribution and (*iv*) $E(X \mid Y = 1)$.

(*e*) Estimate X when $Y = 10$, if the two lines of regression are :

$$X = -\frac{1}{18} Y + \lambda \text{ and } Y = -2x + \mu,$$

(λ, μ) being unknown and the mean of the distribution is at $(-1, 2)$. Also compute r, λ and μ. [*Gujarat Univ. B.Sc., Oct. 1992*]

11. (*a*) The following results were obtained in the analysis of data on yield of dry bark in ounces (Y) and age in years (X) of 200 cinchona plants :

	X	Y
Average	9·2	16·5
Standard deviation	2·1	4·2

Correlation coefficient = +0·84

Construct the two lines of regression and estimate the yield of dry bark of a plant of age 8 years. [*Patna Univ. B.Sc., 1991*]

(*b*) The following data pertain to the marks in subjects A and B in a certain examination :

Mean marks in A = 39·5

Mean marks in B = 47·5

Standard deviation of marks in A = 10·8

Standard deviation of marks in B = 16·8

Coefficient of correlation between marks in A and marks in B = 0·42.

Draw the two lines of regression and explain why there are two regression equations. Give the estimate of marks in B for candidates who secured 50 marks in A.

Ans. $Y = 0{\cdot}65X + 21{\cdot}825$, $X = 0{\cdot}27Y + 26{\cdot}675$ and $Y = 54{\cdot}342$ for $X = 50$

(c) You are given the following information about advertising expenditure and sales :

	Advertising Expenditure (X) *(Rs. lakhs)*	*Sales (Y)* *(Rs. lakhs)*
Mean	10	90
s.d.	3	12

Correlation coefficient = 0·8

What should be the advertising budget if the company wants to attain sales target of Rs. 120 lakhs ? [*Delhi Univ. M.C.A., 1990*]

12. Twenty-five pairs of value of variates X and Y led to the following results :

$N = 25, \Sigma X = 127, \Sigma Y = 100, \Sigma X^2 = 760, \Sigma Y^2 = 449$ and $\Sigma XY = 500$

A subsequent scrutiny showed that two pairs of values were copied down as :

X	Y
8	14
8	6

X	Y
8	12
6	8

(i) Obtain the correct value of the correlation coefficient.

(ii) Hence or otherwise, find the correct equations of the two lines of regression.

(iii) Find the angle between the regression lines.

Ans. (i) $r(X, Y) = -(0{\cdot}64 \times 0{\cdot}15)^{1/2}$,

(ii) $X = -0{\cdot}64Y + 7{\cdot}56$, $Y = -0{\cdot}15X + 4{\cdot}75$.

13. Suppose you have n observations :

$$(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$$

on two variables X and Y, and you have fitted a linear regression $Y = a + bX$ by the method of least squares. Denote the 'expected' value of Y by Y^*, and the residual $Y - Y^*$ by e. Find means and variances of Y^* and e, and the correlation co-efficient between (i) X and e, (ii) Y and e and (iii) Y and Y^*. Use these results to bring out the significance and limitations of the correlation coefficient.

Ans. $r(X, e) = 0, r(Y, e) = 0$ and $r(Y, Y^*) = r(X, Y)$.

14. (a) The regression lines of Y on X and of X on Y are respectively $Y = aX + b$ and $X = cY + d$. Show that

(i) Means are $\bar{X} = (bc + d)/(1 - ac)$ and $\bar{Y} = (ad + b)/(1 - ac)$

(ii) Correlation coefficient between X and Y is $\sqrt{a\,c}$

(iii) The ratio of the standard deviations of X and Y is $\sqrt{c/a}$.

(b) For two random variables X and Y with the same mean, the two regression equations are $Y = aX + b$ and $X = \alpha Y + \beta$. Show that $\dfrac{b}{\beta} = \dfrac{1 - a}{1 - \alpha}$.

Find also the common mean.

[*Punjab Univ. B.Sc. (Maths Hons.), 1992*]

(c) If the lines of regression of Y on X and X on Y are respectively $a_1X + b_1Y + c_1 = 0$ and $a_2X + b_2Y + c_2 = 0$, prove that $a_1b_2 \le a_2b_1$.

(Delhi Univ. B.Sc. (Stat. Hons.), 1989)

Hint. $r^2 = b_{YX} \cdot b_{XY} \le 1 \Rightarrow \left(-\dfrac{a_1}{b_1}\right) \times \left(-\dfrac{b_2}{a_2}\right) = \dfrac{a_1b_2}{a_2b_1} \le 1$

15. (a) By minimising $\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)^2$ for variations in α and p, show that there are two straight lines passing through the mean of the distribution for which the sum of squares of normal deviations has an extreme value. Prove also that their slopes are given by

$$\tan 2\alpha = \frac{2\mu_{11}}{\sigma_X^2 - \sigma_Y^2}$$

Hint. We have to minimize

$$S = \sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)^2 \qquad \ldots(1)$$

Equating to zero, the partial derivatives of (1) w.r.t. α and p, we have

$$\frac{\partial S}{\partial \alpha} = 0 = 2 \sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)(-x_i \sin\alpha + y_i \cos\alpha) \qquad \ldots(2)$$

$$\frac{\partial S}{\partial p} = 0 = -2 \sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p) \qquad \ldots(3)$$

Equation (3) can be written as

$$\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p) = 0 \Rightarrow \bar{x} \cos\alpha + \bar{y} \sin\alpha - p = 0 \qquad \ldots(4)$$

From equation (2), we get a quadratic equation which shows that there are two straight lines for extreme values of E.

From equation (4), it becomes clear that both the straight lines pass through the point $(\bar{x}, \bar{y})$.

Again equation (2) can be written as :

$$\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)(y_i \cos\alpha - x_i \sin\alpha) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} f_i [\cos\alpha (x_i - \bar{x}) + \sin\alpha (y_i - \bar{y})][y_i \cos\alpha - x_i \sin\alpha] = 0$$

[Using (4)]

$$\Rightarrow \cos^2\alpha \sum_{i=1}^{n} f_i y_i (x_i - \bar{x}) - \sin\alpha \cos\alpha \sum_{i=1}^{n} f_i x_i (x_i - \bar{x})$$

$$+ \sin\alpha \cos\alpha \sum_{i=1}^{n} f_i y_i (y_i - \bar{y}) - \sin^2\alpha \sum_{i=1}^{n} f_i x_i (y_i - \bar{y}) = 0 \qquad \ldots(5)$$

We have $\quad \mu_{11} = \dfrac{1}{N} \sum_i f_i (x_i - \bar{x})(y_i - \bar{y})$

$$= \frac{1}{N} \sum_i f_i x_i (y_i - \bar{y}) - \bar{x} . \frac{1}{N} \sum_i f_i (y_i - \bar{y}) = \frac{1}{N} \sum f_i x_i (y_i - \bar{y})$$

Similarly, $\mu_{11} = \frac{1}{N} \sum_i f_i y_i (x_i - \bar{x})$

$$\sigma_X^2 = \frac{1}{N} \sum_i f_i x_i (x_i - \bar{x}) \text{ and } \sigma_Y^2 = \frac{1}{N} \sum_i f_i y_i (y_i - \bar{y})$$

Substituting these values in (5), we get the required result.

(b) If the straight line defined by

$$Y = a + bX$$

satisfies the condition $E[(Y - a - bX)^2]$ = minimum, show that the regression line of the random variable Y on the random variable X is

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X}), \text{ where } \bar{X} = E(X), \bar{Y} = E(Y)$$

16. (a) Define Curve of regression of Y on X.

The joint density function of X and Y is given by :

$$f(x, y) = x + y , 0 < x < 1, 0 < y < 1$$
$$= 0, \text{ otherwise}$$

Find

(i) the correlation coefficient between X and Y,

(ii) the regression curve of Y on X, and

(iii) the regression curve of X on Y.

Ans. $\rho (X, Y) = -\frac{1}{11}$. *[Madras Univ. B.Sc., Stat. (Main), 1992]*

(b) Let $f(x_1, x_2) = \frac{2}{a^2} ; 0 < x_1 < x_2, 0 < x_2 < a$

$$= 0, \text{ elsewhere}$$

be the joint p.d.f. of X_1 and X_2.

Find conditional means and variances. Also show that $\rho = \frac{1}{2}$.

17. If the joint density of X and Y is given by

$$f(x, y) = \begin{cases} (x + y)/3, & \text{for } 0 < x < 1, 0 < y < 2 \\ 0, & \text{otherwise} \end{cases}$$

obtain the regressions (i) of Y on X and (ii) of X on Y.

Are the regressions linear ? Find the correlation coefficient between X and Y. *(Allahabad Univ. B.Sc. 1992)*

Ans. $y = E(Y \mid x) = \frac{3x + 4}{3(x + 1)} ; x = E(X \mid y) = \frac{2 + 3y}{3(1 + 2y)}$

$$\text{Corr. } (X, Y) = -\left(\frac{2}{299} \right)^{1/2}$$

18. Let the joint density function of X and Y be given by :

$$f(x, y) = 8xy , 0 < x < y < 1$$
$$= 0 , \text{ otherwise}$$

Find: (i) $E(Y \mid X = x)$, (ii) $E[XY \mid X = x]$, (iii) $\text{Var}\,[Y \mid X = x]$

[Delhi Univ. BSc. (Maths Hons.), 1988]

Ans. (i) $E(Y|x) = \frac{2}{3}\left(\frac{1+x+x^2}{1+x}\right)$, $E\,(XY \mid x) = x\,E(Y \mid x)$, (iii) $E\,(Y^2 \mid x) = \frac{1+x^2}{2}$

19. Give an example to show that it is possible to have the regression of Y on X constant (does not depend on X), but the regression of X on Y is not constant (does depend on Y).

Hint. See Example 10·21.

20. Prove or disprove :

$$E(Y \mid X = x) = \text{constant} \;\Rightarrow\; r(X,Y) = 0$$

Ans. True

21. If $f(x,y) = \frac{1}{3}x^2 \exp\,[-y\,(1+x)]$, $x \geq 0, y \geq 0$ is the joint p.d.f. of (X,Y), obtain the equation of regression of Y on X.

Ans. $y = E(Y \mid x) = 1/(1+x)$.

22. Variables (X,Y) have joint p.d.f.

$f(x,y) = 6(1-x-y),\ x > 0,\ y > 0,\ x + y < 1.$

$= 0$, otherwise.

Find $f_X(x), f_Y(y)$ and Cov (X,Y). Are X and Y independent ? Obtain the regression curves for the means.

[Calcutta Univ. B.Sc. (Maths Hons.), 1986]

Ans. $f_1(x) = 3(1-x)^2,\ 0 < x < 1$; $f_2(y) = 3(1-y)^2,\ 0 < y < 1$.

X and Y are not independent.

Regression curves for the means are :

$$y = E(Y \mid x) = \tfrac{1}{3}(1-x) \text{ and } x = E(X \mid y) = \tfrac{1}{3}(1-y).$$

23. For the joint p.d.f.

$$f(x, y) = 3x^2 - 8xy + 6y^2;\ 0 \leq (x, y) \leq 1,$$

find the least square regression lines and the regression curves for the means.

[Calcutta Univ. B.Sc. (Maths. Hons.), 1987]

Ans. Regression lines :

$$y - \frac{2}{3} = -\frac{10}{67}\left(x - \frac{5}{12}\right) \;;\quad x - \frac{5}{12} = -\frac{25}{32}\left(y - \frac{2}{3}\right)$$

Regression curves for means are :

$$y = E(Y|x) = \frac{9x^2 - 16x + 9}{6(3x^2 - 4x + 2)} \;;\quad x = E(X \mid y) = \frac{36y^2 - 32y - 9}{12(6y^2 - 4y + 1)}$$

24. Let (X, Y) be jointly distributed with p.d.f.

$f(x, y) = e^{-y},\ 0 < x < y < \infty$

$= 0$, otherwise

Prove that :

$$E(Y \mid X = x) = x + 1 \quad \text{and} \quad E(X \mid Y = y) = y/2.$$

Hence prove that $r(X, Y) = \sqrt{1/2}$.

25. Let $f(x, y) = e^{-y}(1 - e^{-x}), 0 < x < y ; 0 < y < \infty$
$= e^{-x}(1 - e^{-y}), 0 < y < x ; 0 < x < \infty$

(a) Show that $f(x, y)$ is a p.d.f.
(b) Find marginal distributions of X and Y.
(c) Find $E(Y \mid X = x)$ for $x > 0$.
(d) Find $P(X \leq 2, Y \leq 2)$.
(e) Find the correlation coefficient $r(X, Y)$.
(f) Find another joint p.d.f. having the same marginals.

Ans. (b) $f_1(x) = xe^{-x}, 0 < x < \infty ; f_2(y) = ye^{-y}, 0 < y < \infty$.

(c) $E(Y \mid x) = \dfrac{1 - e^{x}}{x}[x - 1] + \dfrac{1}{x}\left(\dfrac{x^2}{2} + xe^{x} + e^{-x} - 1\right)$

(d) $1 - \dfrac{1}{e^4} - \dfrac{4}{e^2}$; (e) $r(x, y) = \dfrac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \dfrac{1}{\sqrt{2}\sqrt{2}} = \dfrac{1}{2}$

(f) **Hint.** $f(x, y, \alpha) = f_1(x) f_2(y)[1 + \alpha(2F(x) - 1)(2F(y) - 1)]$, $|\alpha| < 1$, has the same marginals $f_1(x)$ and $f_2(y)$.

26. Obtain regression equation of Y on X for the distributions :

(a) $f(x, y) = \dfrac{9}{2} \cdot \dfrac{1 + x + y}{(1 + x)^4 (1 + y)^4}$; $x, y \geq 0$

(b) $f(x, y) = \dfrac{4}{5}(x + 3y)e^{-x-2y}$; $x, y \geq 0$

[*Sardar Patel Univ. M.Sc., 1992*]

Ans. (a) **Hint.** See Example 5·25, page 5·55, (b) $\dfrac{x + 3}{2x + 3}$.

27. A ball is drawn at random from an urn containing three white balls numbered 0, 1, 2 ; two red balls numbered 0, 1 and one black ball numbered 0. If the colours white, red and black are again numbered 0, 1 and 2 respectively, find the correlation coefficient between the variates X, the colour number and Y the number of the ball. Write down the equation of regression line of Y on X.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

OBJECTIVE TYPE QUESTIONS

I. State, giving reasons, whether each of the following statements is true or false.

(i) Both regression lines of Y on X and of X on Y do not intersect at all.
(ii) In a bivariate regression, $b_{YX} = \frac{1}{5}$, $b_{XY} = 10$
(iii) The regression coefficient of Y on X is 3·2 and that of X on Y is 0·8.
(iv) There is no relationship between correlation coefficient and regression coefficient.
(v) Both the regression coefficients cannot exceed unity.

(*vi*) The greater the value of 'r', the better are the estimates obtained through regression analysis.

(*vii*) If X and Y are negatively correlated variables, and (0, 0) is on the least squares line of Y on X, and if $X = 1$ is the observed value then predicted value of Y must be negative.

(*viii*) Let the correlation between X and Y be perfect and positive. Suppose the points (3, 5) and (1, 4) are on the regression lines. With this knowledge it is possible to determine the least squares line exactly.

(*ix*) If the lines of regression are $Y = \frac{1}{4}X$ and $X = \frac{1}{9}Y + 1$, then $\rho = \frac{1}{6}$ and $E(X \mid Y = 0) = 1$.

(*x*) In a bivariate distribution, $b_{YX} = 2{\cdot}8$ and $b_{XY} = 0{\cdot}3$.

II. Fill in the blanks :

(*i*) The regression analysis measures ... between X and Y.

(*ii*) Lines of regression are ... if $r_{XY} = 0$ and they are ... if $r_{XY} = \pm 1$.

(*iii*) If the regression coefficients of X on Y and Y on X are $-0{\cdot}4$ and $-0{\cdot}9$ respectively then the correlation coefficient between X and Y is ...

(*iv*) If the two regression lines are $X + 3Y - 5 = 0$ and $4X + 3Y - 8 = 0$, then the correlation coefficient between X and Y is ...

(*v*) If one of the regression coefficients is ... unity, the other must be ... unity.

(*vi*) The farther the two regression lines cut each other, the ... will be the degree of correlation.

(*vii*) When one regression coefficient is positive, the other would be ...

(*viii*) The sign of regression coefficient is ... as that of correlation coefficient.

(*ix*) Correlation coefficient is the ... between regression coefficients.

(*x*) Arithmetic mean of regression coefficients is ... correlation coefficient.

(*xi*) When the correlation coefficient is zero, the two regression lines are ... and when it is ± 1, then the regression lines are ...

III. Indicate the correct answer :

(*i*) The regression line of Y on X (*a*) minimises total of the squares of horizontal deviations, (*b*) total of the squares of the vertical deviations, (*c*) both vertical and horizontal deviations, (*d*) none of these.

(*ii*) The regression coefficients are b_2 and b_1. Then the correlation coefficient r is (*a*) b_1/b_2, (*b*) b_2/b_1, (*c*) $b_1 b_2$ (*d*) $\pm\sqrt{b_1 b_2}$.

(*iii*) The farther the two regression lines cut each other (*a*) the greater will be the degree of correlation, (*b*) the lesser will be the degree of correlation, (*c*) does not matter.

(iv) If one regression coefficient is greater than unity, then the other must be (a) greater than the first one, (b) equal to unity, (c) less than unity, (d) equal to zero.

(v) When the correlation coefficient $r = \pm 1$, then the two regression lines (a) are perpendicular to each other, (b) coincide, (c) are parallel to each other, (d) do not exist.

(vi) The two lines of regression are given as $X + 2Y - 5 = 0$ and $2X + 3Y = 8$. Then the mean values of X and Y respectively are (a) 2, 1, (b) 1, 2, (c) 2, 5, (d) 2, 3.

(vii) The tangent of the angle between two regression lines is given as 0·6 and the s.d. of Y is known to be twice that of X. Then the value of correlation coefficient between X and Y is (a) $-\frac{1}{2}$, (b) $\frac{1}{2}$, (c) 0·7, (d) 0·3.

IV. σ_X and σ_Y are the standard deviations of two correlated variables X an Y respectively in a large sample, and r is the sample correlation coefficient.

(i) State the "Standard Error of Estimate" for linear regression of Y on X.

(ii) What is the standard error in estimating Y from X if $r = 0$?

(iii) By how much is this error reduced if r is increased to 0·30 ?

(iv) How large must r be in order to reduce this standard error to one-half its value for $r = 0$?

(v) Give your interpretations for the cases $r = 0$ and $r = 1$.

V. Explain why we have two lines of regression.

10·8. Correlation Ratio. As discused earlier, when variables are linearly related, we have the regression lines of one variable on another variable and correlation coefficient can be computed to tell us about the extent of association between them. However, if the variables are not linearly related but some sort of curvilinear relationship exists between them, the use of r which is a measure of the degree to which the relation approaches a straight line "law" will be misleading. We might come across bivariate distributions where r may be very low or even zero but the regression may be strong, or even perfect. *Correlation ratio* 'η' is the appropriate measure of curvilinear relationship between the two variables. Just as r measures the concentration of points about the straight line of best fit, η measures the concentration of points about the curve of best fit. If regression is linear $\eta = r$, otherwise $\eta > r$ (c.f. Remark 2, § 10·8·1).

10·8·1. Measure of Correlation Ratio. In the previous articles we have assumed that there is a single observed value Y corresponding to the given value x_i of X but sometimes there are more than one such value of Y.

Suppose corresponding to the values x_i, $(i = 1, 2, \ldots, m)$ of the variable X, the variable Y takes the values y_{ij} with respective frequencies f_{ij}, $j = 1, 2, \ldots, n$.

Though all the x's in the ith vertical array have the same value, the y's are different. A typical pair of values in the ith array is (x_i, y_{ij}), with frequency f_{ij}.

Thus the first suffix i indicates the vertical array while the second suffix j indicates the positions of y in that array. Let

$$\sum_{j=1}^{n} f_{ij} = n_i \quad \text{and} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} f_{ij}\right) = \sum_{i=1}^{m} n_i = N, \quad \text{(say)}.$$

If $\bar{y}_i$ and $\bar{y}$ denote the means of the ith array and the overall mean respectively, then

$$\bar{y}_i = \frac{\sum_j f_{ij}\, y_{ij}}{\sum_j f_{ij}} = \frac{\sum_j f_{ij}\, y_{ij}}{n_i} = \frac{T_i}{n_i} \quad \text{and} \quad \bar{y} = \frac{\sum_i\sum_j f_{ij}\, y_{ij}}{\sum_i\sum_j f_{ij}} = \frac{\sum_i n_i\, \bar{y}_i}{\sum_i n_i} = \frac{T}{N}$$

In other words $\bar{y}$ is the weighted mean of all the array means, the weights being the array frequencies.

Def. *The correlation ratio of Y on X, usually denoted by η_{YX} is given by*

$$\eta_{YX}^2 = 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2} \qquad \text{...(10·21)}$$

where σ_{eY}^2 and σ_Y^2 are given by

$$\sigma_{eY}^2 = \frac{1}{N}\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 \quad \text{and} \quad \sigma_Y^2 = \frac{1}{N}\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y})^2$$

A convenient expression for η_{YX} can be obtained in terms of standard deviation σ_{mY} of the means of the vertical arrays, each mean being weighted by the array frequency.

We have

$$\begin{aligned} N\sigma_Y^2 &= \sum\sum f_{ij}\,(y_{ij} - \bar{y})^2 = \sum_i\sum_j f_{ij}\,\{(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})\}^2 \\ &= \sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 + \sum_i\sum_j f_{ij}\,(\bar{y}_i - \bar{y})^2 + 2\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) \end{aligned}$$

The term $2[\sum_i(\bar{y}_i - \bar{y})\,\{\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)\}]$ vanishes since $\sum_j f_{ij}\,(y_{ij} - \bar{y}_i) = 0$, being the algebraic sum of the deviations from mean.

$$\begin{aligned} \therefore \qquad N\sigma_Y^2 &= \sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 + \sum_i n_i(\bar{y}_i - \bar{y})^2 \\ \Rightarrow \qquad N\sigma_Y^2 &= N\sigma_{eY}^2 + N\sigma_{mY}^2 \;\Rightarrow\; \sigma_Y^2 = \sigma_{eY}^2 + \sigma_{mY}^2 \\ \Rightarrow \qquad 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2} &= \frac{\sigma_{mY}^2}{\sigma_Y^2} \end{aligned}$$

which on comparison with (10·21) gives

$$\eta_{YX}^2 = \frac{\sigma_{mY}^2}{\sigma_Y^2} = \frac{\sum_i n_i(\bar{y}_i - \bar{y})^2}{\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y})^2} \qquad \text{...(10·22)}$$

We have

$$N\sigma_{m_Y}^2 = \sum_i n_i (\bar{y}_i - \bar{y})^2 = \sum_i n_i \bar{y}_i^2 - N\bar{y}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$$\therefore \qquad \eta_{YX}^2 = \left[\sum_i \left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] \Big/ N\sigma_Y^2, \qquad \ldots(10\cdot23)$$

a formula, much more convenient for computational purposes.

Remarks 1. (10·21) implies that

$$\sigma_{eY}^2 = \sigma_Y^2 (1 - \eta_{YX}^2)$$

Since σ_{eY}^2 and σ_Y^2 are non-negative, we have

$$1 - \eta_{YX}^2 \geq 0 \quad \Rightarrow \quad \eta_{YX}^2 \leq 1 \quad \Rightarrow \quad |\eta_{YX}| \leq 1$$

2. Since the sum of squares of deviations in any array is minimum when measured from its mean, we have

$$\sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 \leq \sum_i \sum_j f_{ij} (y_{ij} - \hat{y}_{ij})^2 \qquad \ldots(*)$$

where $\hat{y}_{ij}$ is the estimate of y_{ij} for given value of $X = x_i$, say, as given by the line of regression of Y on X *i.e.*, $\hat{y}_{ij} = a + bx_i$, $(j = 1, 2, \ldots, n)$.

$$\text{But} \qquad \sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 = N\sigma_{eY}^2 = N\sigma_Y^2 (1 - \eta_{YX}^2)$$

$$\text{and} \qquad \sum_i \sum_j f_{ij} (y_{ij} - a - bx_i)^2 = N\sigma_Y^2 (1 - r^2) \qquad (c.f.\ \S\ 10\cdot7\cdot6)$$

$$\therefore\ (*) \Rightarrow \qquad 1 - \eta_{YX}^2 \leq 1 - r^2$$

$$i.e., \qquad \eta_{YX}^2 \geq r^2 \quad \Rightarrow \quad |\eta_{YX}| \geq |r|$$

Thus the absolute value of the correlation ratio can never be less than the absolute of r, the correlation coefficient.

When the regression of Y on X is linear, straight line of means of arrays coincides with the line of regression and $\eta_{YX}^2 = r^2$. Thus $\eta_{YX}^2 - r^2$ is the departure of regression from linearity. It is also clear (from Remark 1) that the more nearly η_{YX}^2 approaches unity, the smaller is σ_{eY}^2 and, therefore, closer are the points to the curve of means of vertical arrays.

$$\text{When} \qquad \eta_{YX}^2 = 1, \sigma_{eY}^2 = 0 \Rightarrow \sum \sum f_{ij} (y_{ij} - \bar{y}_i)^2 = 0$$

$\Rightarrow y_{ij} = \bar{y}_i$, $\forall\ j = 1, 2, \ldots, n$, *i.e.*, all the points lie on the curve of means. This implies that there is a functional relationship between X and Y. η_{YX} is, therefore, the measure of the degree to which the association between the variables approaches a functional relationship of the form $Y = F(X)$, where $F(X)$ is a single valued function of X, $[F(X) = a + bX]$.

3. It is worth noting that the value of η_{YX} is not independent of the classification of the data. As the class intervals become narrower η_{YX} approaches unity, since in that case σ_{mY}^2 gets nearer to σ_Y^2. If the grouping is so fine that only one item appears in each row (related to each x-class), that item will constitute the mean of that column and thus in this case σ_{mY}^2 and σ_Y^2 become equal so that $\eta_{YX}^2 = 1$. On the other hand, a very coarse grouping tends to make the value of η_{YX} approach r. "Student" has given a formula for 'the correction' to be made in the correlation ratio 'for grouping' in Biometrika (Vol IX page 316-320.)

4. It can be easily proved that η_{YX}^2 is independent of change of origin and scale of measurements.

5. η_{XY}^2, the second correlation ratio of X on Y depends upon the scatter of observations about the line of column means.

6. r_{XY} and r_{YX} are same but η_{YX} is, in general, different from η_{XY}.

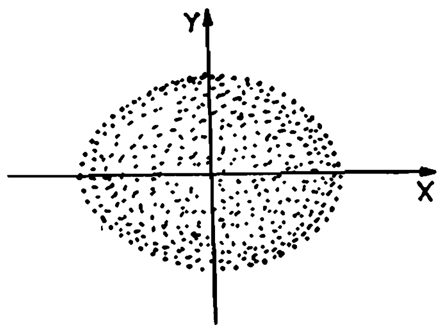
7. In terms of expectation, correlation ratio is defined as follows :

$$\eta_{YX}^2 = \frac{E_X\,[E(Y\,|\,X) - E(Y)]^2}{E[Y - E(Y)]^2} = \frac{E[E(Y\,|\,X) - E(Y)]^2}{\sigma_Y^2}.$$

and
$$\eta_{XY}^2 = \frac{E_Y\,[E\,(X\,|\,Y) - E(X)]^2}{E[X - E(X)]^2} = \frac{E[E(X\,|\,Y) - E(X)]^2}{\sigma_X^2}$$

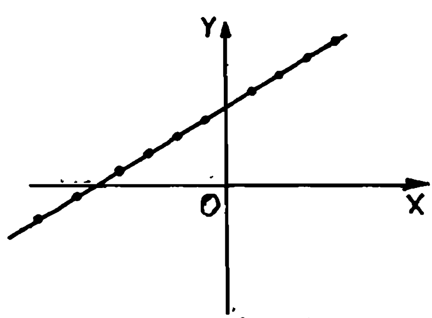
8. We give below some diagrams, exhibiting the relationship between r and η_{YX}.

(i) For completely random scattering of the dots with no trend, both r and η are zero.



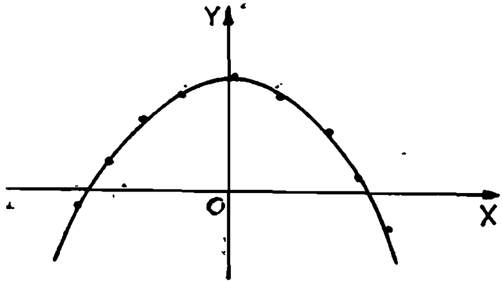
$r = 0, \eta_{YX} = \eta_{XY} = 0$

(ii) If dots lie precisely on a line, $r = 1$ and $\eta = 1$.



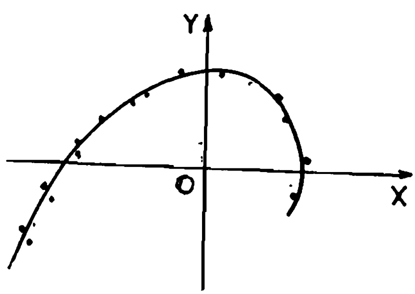
$r = 1, \eta_{YX} = \eta_{XY} = 1$

(iii) If dots lie on a curve, such that no ordinate cuts it more than once, $\eta_{YX} = 1$ and if furthermore, the dots are symmetrically placed about Y-axis, then $\eta_{XY} = 0$, $r = 0$.



$r = 0$, $\eta_{YX} = 1$, $\eta_{XY} = 0$

(iv) If $\eta_{YX} > r$, the dots are scattered around a definitely curved trend line.



$\eta_{YX} > r$

EXERCISE 10(e)

1. (*a*) Define correlation coefficient and correlation ratio. When is the latter a more suitable measure of correlation than the former ? Show that the correlation ratio is never less than the correlation coefficient. What do you infer if the two are equal ? Further, show that none of these can exceed one.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

(*b*) Show that $$1 \geq \eta_{YX}^2 \geq r_{YX}^2 \geq 0$$

Interpret each of the following statements.

(*i*) $r = 0$, (*ii*) $r^2 = 1$, (*iii*) $\eta^2 = 1$, (*iv*) $\eta^2 = r^2$ and (*v*) $\eta = 0$

(*c*) When the correlation coefficient is equal to unity, show that the two correlation ratios are also equal to unity. Is the converse true ?

(*d*) Define correlation ratio η_{XY} and prove that

$$1 \geq \eta^2_{XY} \geq r^2,$$

where r is the coefficient of correlation between X and Y. Show further that $(\eta^2_{XY} - r^2)$ is a measure of non-linearity of regression.

2. For the joint p.d.f.

$$f(x, y) = \frac{1}{2} x^3 \exp[-x(y+1)], \; y > 0, x > 0$$
$$= 0 \quad , \text{ otherwise,}$$

find :

(i) Two lines of regression.

(ii) The regression curves for the means.

(iii) $r(X, Y)$.

(iv) η^2_{YX} and η^2_{XY}.

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1987*]

Ans. (i) $y = -\frac{1}{6}x + 1 \quad ; \quad x = -\frac{2}{3}y + \frac{10}{3}$

(ii) $y = E(Y \mid x) = \dfrac{1}{x} \quad ; \quad x = E(X \mid y) = \dfrac{4}{1+y}$

(iii) $r(X, Y) = -\frac{1}{3}$ (iv) $\eta^2_{YX} = \frac{1}{3}, \eta^2_{XY} = \frac{1}{5}$

3. Compute $r(X, Y)$ and η_{YX} for the following data :

X :	0·5 — 1·5	1·5 — 2·5	2·5 — 3·5	3·5 — 4·5	4·5 — 5·5
f :	20	30	35	25	15
$\bar{y}_i$:	11·3	12·7	14·7	16·5	19·1

Var $(Y) = 9·61$

Ans. $\eta_{YX} = 0·77, \; r = 0·85$

4. Compute η_{XY} for the following table :

$Y \downarrow$ \ $X \rightarrow$	47	52	57	62	67
57	4	4	2	...	...
62	4	8	8	1	...
67	...	7	12	1	4
72	...	3	1	8	5
77	...	...	3	5	6

10·9. Intra-class Correlation. Intra-class correlation means within class correlation. It is distinguishable from product moment correlation in as much as here both the variables measure the same characteristics. Sometimes specially in biological and agricultural study, it is of interest to know how the members of a family or group are correlated among themselves with respect to some one of their common characteristic. For example, we may require the correlation between the heights of brothers of a family or between yields of plots of an experimental block. In such cases both the variables measure the same characteristic, *e.g.*, height and height or weight and weight. There is nothing to distinguish one from the other so that one may be treated as X-variable and the other as the Y-variable.

Suppose we have $A_1, A_2, \ldots, A_n$ families with $k_1, k_2, \ldots, k_n$ members, each of which may be represented as

$$\begin{array}{cccc}
x_{11} & x_{21} \cdots\cdots & x_{i1} \cdots\cdots & x_{n1} \\
x_{12} & x_{22} & x_{i2} & x_{n2} \\
\vdots & \vdots & \vdots & \vdots \\
x_{1j} & x_{2j} \cdots\cdots & x_{ij} \cdots\cdots & x_{nj} \\
\vdots & \vdots & \vdots & \vdots \\
x_{1k_1} & x_{2k_2} \cdots\cdots & x_{ik_i} \cdots\cdots & x_{nk_n}
\end{array}$$

and let x_{ij} $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, k_i)$ denote the measurement on the jth member in the ith family.

We shall have $k_i(k_i - 1)$ pairs for the ith family or group like (x_{ij}, x_{il}), $j \neq l$. There will be $\sum\limits_{i=1}^{n} k_i (k_i - 1) = N$ pairs for all the n families or groups. If we prepare a correlation table there will be $k_i (k_i - 1)$ entries for the ith group or family and $\sum\limits_i k_i (k_i - 1) = N$ entries for all the n families or groups. The table is symmetrical about the principal diagonal. Such a table is called an *intra-class correlation table* and the correlation is called *intra-class correlation*.

In the bivariate table x_{i1} occurs $(k_i - 1)$ times, x_{i2} occurs $(k_i - 1)$ times, ..., x_{ik_i} occurs $(k_i - 1)$ times, *i.e.*, from the ith family we have $(k_i - 1) \sum\limits_j x_{ij}$ and hence for all the n families we have $\sum\limits_i (k_i - 1) \sum x_{ij}$ as the marginal frequency, the table being symmetrical about principal diagonal.

$$\therefore \qquad \bar{x} = \bar{y} = \frac{1}{N}\left[\sum_i (k_i - 1) \sum_j x_{ij}\right]$$

Similarly,

$$\sigma_X^2 = \sigma_Y^2 = \frac{1}{N}\left[\sum_i (k_i - 1) \sum_j (x_{ij} - \bar{x})^2\right]$$

Further

$$\text{Cov}(X, Y) = \frac{1}{N} \sum_i \left[\sum_{j,\, l} (x_{ij} - \bar{x})(x_{il} - \bar{x})\right], \; j \neq l$$

$$= \frac{1}{N} \sum_i \left[\sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{ij} - \bar{x})(x_{il} - \bar{x}) - \sum_{j=1}^{k_i} (x_{ij} - \bar{x})^2\right]$$

If we write $\bar{x}_i = \sum_j x_{ij} / k_i$, then

$$\sum_i \left[\sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right] = \sum_i \left[\sum_j (x_{ij} - \bar{x}) \sum_l (x_{il} - \bar{x}) \right]$$

$$= \sum_i [k_i (\bar{x}_i - \bar{x}) k_i (\bar{x}_i - \bar{x})]$$

$$= \sum_i k_i^2 (\bar{x}_i - \bar{x})^2$$

Therefore intra-class correlation coefficient is given by

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)\,V(Y)}} = \frac{\sum_i k_i^2 (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{\sum_i \sum_j (k_i - 1)(x_{ij} - \bar{x})^2} \qquad \ldots(10\cdot24)$$

If we put $k_i = k$, *i.e.*, if all families have equal members then

$$r = \frac{k^2 \sum_i (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{(k-1) \sum_i \sum_j (x_{ij} - \bar{x})^2}$$

$$= \frac{nk^2 \sigma_m^2 - nk\sigma^2}{(k-1)\, nk\sigma^2} = \frac{1}{(k-1)} \left\{ \frac{k\,\sigma_m^2}{\sigma^2} - 1 \right\} \qquad \ldots(10\cdot24a)$$

where σ^2 denotes the variance of X and σ_m^2 the variance of means of families.

Limits. We have from (10·24a),

$$1 + (k-1)\, r = \frac{k\sigma_m^2}{\sigma^2} \geq 0 \quad \Rightarrow \quad r \geq -\frac{1}{(k-1)}$$

Also $\quad 1 + (k-1)\, r \leq k$, as the ratio $\dfrac{\sigma_m^2}{\sigma^2} \leq 1 \quad \Rightarrow \quad r \leq 1$

so that $\quad -\dfrac{1}{(k-1)} \leq r \leq 1$

Interpretation. Intraclass correlation cannot be less than $-1/(k-1)$, though it may attain the value +1 on the positive side, so that it is a skew coefficient and a negative value has not the same significance as a departure from independence as an equivalent positive value.

EXERCISE 10 (f)

1. If $x_1, x_2, \ldots, x_k$ be k variates with standard deviation σ and m be any number, prove that

$$k^2\sigma^2 = (k-1) \sum_{r=1}^{k} (x_r - m)^2 - \sum_{r=1}^{k} \sum_{s=1}^{k} (x_r - m)(x_s - m), \; r \neq s$$

Hence deduce that the coefficient of intraclass correlation for n families with varying number of members in each family is

$$1 - \frac{\sum_i k_i \sigma_i^2}{\sigma^2 \sum_i k_i(k_i - 1)}$$

where k_i, σ_i^2 denote the number of members and the variance respectively in the *i*th family and σ^2 is the general variance.

Given $n = 5$, $\sigma_i = i$, $k_i = i + 1$ $(i \le 5)$, find the least possible intraclass correlation coefficient.

2. What do you understand by intra-class correlation coefficient.

Calculate its value for the following data :

Family No.	*Height of brothers*			
1	60	62	63	65
2	59	60	61	62
3	62	62	64	63
4	65	66	65	66
5	66	67	67	69

3. In four families each containing eight persons, the chest measurements of persons are given below. Calculate the intraclass correlation co-efficient.

Family	1	2	3	4	5	6	7	8
I	43	46	48	42	50	45	45	49
II	33	34	37	39	82	35	37	41
III	56	52	50	51	54	52	39	52
IV	34	37	38	40	40	41	44	44

10·10. Bivariate Normal Distribution. The bivariate normal distribution is a generalization of a normal distribution for a single variate. Let X and Y be two normally correlated variables with correlation coefficient ρ and $E(X) = \mu_1$, Var $(X) = \sigma_1^2$; $E(Y) = \mu_2$, Var $(Y) = \sigma_2^2$. In deriving the bivariate normal distribution we make the following three assumptions.

(*i*) *The regression of Y on X is linear.* Since the mean of each array is on the line of regression $Y = \rho(\sigma_2/\sigma_1)X$, the mean or expected value of Y is $\rho(\sigma_2/\sigma_1)X$, for different values of X.

(*ii*) *The arrays are homoscedastic, i.e.,* variance in each array is same. The common variance of estimate of Y in each array is then given by $\sigma_2^2(1 - \rho^2)$, ρ being the correlation coefficient between variables X and Y and is independent of X.

(*iii*) *The distribution of Y in different arrays in normal.* Suppose that one of the variates, say X, is distributed normally with mean 0 and standard deviation σ_1 so that the probability that a random value of X will fall in the small interval dx is

$$g(x)\,dx = \frac{1}{\sigma_1\sqrt{(2\pi)}} \cdot \exp\,(-x^2/2\sigma_1^2)\,dx$$

The probability that a value of Y, taken at random in an assigned vertical array will fall in the interval dy is

$$h(y \mid x)\, dy = \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(y - \rho\, x\, \frac{\sigma_2}{\sigma_1}\right)^2\right\}$$

The joint probability differential of X and Y is given by

$$dP(x, y) = g(x) h(y \mid x) dx dy$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot e^{-\frac{1}{2\sigma_1^2}\cdot x^2}\; e^{\left[-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(y - \rho\frac{\sigma_2}{\sigma_1}x\right)^2\right]}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)\right\}$$

Shifting the origin to (μ_1, μ_2), we get

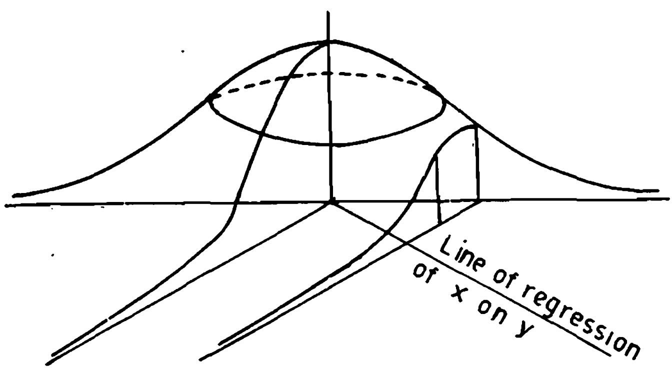
$f_{XY}(x, y)$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}};$$

$$(-\infty < x < \infty, -\infty < y < \infty) \quad \ldots(10\cdot25)$$

where $\mu_1, \mu_2, \sigma_1\ (>0), \sigma_2\ (>0)$ and $\rho\ (-1 < \rho < 1)$ are the five parameters of the distribution.

NORMAL CORRELATION SURFACE



This is the density function of a bivariate normal distribution. The variables X and Y are said to be normally correlated and the surface $z = f(x, y)$ is known as the *normal correlation surface*. The nature of the normal correlation surface is indicated in the above diagram

Remarks 1. The vector $(X, Y)'$ following the joint p.d.f. $f(x, y)$ as given in (10·25), will be abbreviated as $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ or $BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. If in particular $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$ then

$$(X, Y) \sim N(0, 0, 1, 1, \rho) \text{ or } BVN(0, 0, 1, 1, \rho).$$

2. The curve $z = f(x, y)$ which is the equation of a surface in three dimensions, is called the '*Normal Correlation Surface*'.

10·10·1. Moment Generating Function of Bivariate Normal Distribution. Let $(X, Y) \sim BVN\ (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. By def.,

$$M_{XY}(t_1, t_2) = E\left[e^{t_1 X + t_2 Y}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(t_1 x + t_2 y).f(x, y)\, dx dy$$

Put $\dfrac{x - \mu_1}{\sigma_1} = u,\ \dfrac{y - \mu_2}{\sigma_2} = v,\ -\infty < (u, v) < \infty$

i.e., $x = \sigma_1 u + \mu_1,\ y = \sigma_2 v + \mu_2 \Rightarrow |J| = \sigma_1\sigma_2$

$$\therefore\ M_{X,Y}(t_1, t_2) = \frac{\exp(t_1\mu_1 + t_2\mu_2)}{2\pi\sqrt{1-\rho^2}}$$

$$\times \iint_{u\ v} \exp\left[t_1\sigma_1 u + t_2\sigma_2 v - \frac{1}{2(1-\rho^2)}\left\{u^2 - 2\rho uv + v^2\right\}\right] dudv$$

$$= \frac{\exp(t_1\mu_1 + t_2\mu_2)}{2\pi\sqrt{1-\rho^2}}$$

$$\times \iint_{u, v} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{(u^2 - 2\rho uv + v^2) - 2(1-\rho^2)(t_1\sigma_1 u + t_2\sigma_2 v)\right\}\right] dudv$$

We have

$$(u^2 - 2\rho uv + v^2) - 2(1-\rho^2)(t_1\sigma_1 u + t_2\sigma_2 v)$$
$$= \{(u - \rho v) - (1-\rho^2)t_1\sigma_1\}^2$$
$$+ (1-\rho^2)\{(v - \rho t_1\sigma_1 - t_2\sigma_2)^2 - t_1^2\sigma_1^2 - t_2^2\sigma_2^2 - 2\rho t_1 t_2\sigma_1\sigma_2\} \quad ..(*)$$

By taking

$$\left.\begin{aligned} u - \rho v - (1-\rho^2)\, t_1\sigma_1 &= \omega(1-\rho^2)^{1/2} \\ \text{and}\quad v - \rho t_1\sigma_1 - t_2\sigma_2 &= z \end{aligned}\right\} \Rightarrow dudv = \sqrt{1-\rho^2}\, dwdz$$

and using (*), we get

$$M_{X,Y}(t_1, t_2) = \exp[t_1\mu_1 + t_2\mu_2 + \tfrac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t_1 t_2\sigma_1\sigma_2)]$$

$$\times \left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\omega^2/2} d\omega\right] \times \left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\, dz\right]$$

$$= \exp[t_1\mu_1 + t_2\mu_2 + \tfrac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t_1 t_2\sigma_1\sigma_2)] \quad \ldots(10·26)$$

In particular if $(X, Y) \sim BVN\ (0, 0, 1, 1, \rho)$, then

$$M_{X,Y}(t_1, t_2) = \exp[\tfrac{1}{2}(t_1^2 + t_2^2 + 2\rho t_1 t_2)] \quad \ldots(10·26a)$$

Theorem 10·5. *Let* $(X, Y) \sim BVN\ (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. *Then X and Y are independent if and only if* $\rho = 0$.

Proof. (*a*) If $(X, Y) \sim BVN\ (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ and $\rho = 0$, then X and Y are independent [*c.f.* Remark 2(*a*) to Theorem 10·2, page 10·5].

Aliter. $(X, Y) \sim BVN\ (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

$\therefore \quad M_{X,Y}(t_1, t_2) = \exp\{t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + 2\rho t_1 t_2\sigma_1\sigma_2 + t_2^2\sigma_2^2)\}$

If $\rho = 0$, then

$$M_{X,Y}(t_1, t_2) = \exp\{t_1\mu_1 + \tfrac{1}{2}t_1^2\sigma_1^2\} \cdot \exp\{t_2\mu_2 + \tfrac{1}{2}t_2^2\sigma_2^2\}$$

$$= M_X(t_1) \cdot M_Y(t_2). \qquad \ldots(*)$$

[$\because$ If $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then the marginal p.d.f.'s of X and Y are normal *i.e.*, $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$].

$(*) \Rightarrow X$ and Y are independent.

(*b*) Conversely if X and Y are independent, then $\rho = 0$ [*c.f.* Theorem 10·2]

Theorem 10·6. *(X, Y) possesses a bivariate normal distribution if and only if every linear combination of X and Y viz., $aX + bY$, $a \neq 0$, $b \neq 0$, is a normal variate.*

Proof. (*a*) Let $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then we shall prove that $aX + bY$, $a \neq 0, b \neq 0$ is a normal variate.

Since (X, Y) has a bivariate normal distribution, we have

$$M_{X,Y}(t_1, t_2) = E(e^{t_1X + t_2Y})$$

$$= e^{t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + 2\rho t_1 t_2\sigma_1\sigma_2 + t_2^2\sigma_2^2)} \qquad \ldots(*)$$

Then m.g.f. of $Z = aX + bY$, is given by :

$$M_Z(t) = E(e^{tZ}) = E(e^{t(aX + bY)}) = E(e^{atX + btY})$$

$$= \exp\{t(a\mu_1 + b\mu_2) + \frac{t^2}{2}(a^2\sigma_1^2 + 2\rho ab\sigma_1\sigma_2 + b^2\sigma_2^2)\},$$

[Taking $t_1 = at$, $t_2 = bt$ in $(*)$]

which is the m.g.f. of normal distribution with parameters

$$\mu = a\mu_1 + b\mu_2\,;\ \sigma^2 = a^2\sigma_1^2 + 2\rho ab\sigma_1\sigma_2 + b^2\sigma_2^2. \qquad \ldots(**)$$

Hence by uniqueness theorem of m.g.f.,

$$Z = aX + bY \sim N(\mu, \sigma^2),$$

where μ and σ^2 are given in $(**)$.

(*b*) Conversely, let $Z = aX + bY$, $a \neq 0$, $b \neq 0$ be a normal variate. Then we have to prove that (X, Y) has a bivariate normal distribution.

Let $\quad Z = aX + bY \sim N(\mu, \sigma^2)$,

where $\quad \mu = EZ = E(aX + bY) = a\mu_x + b\mu_y$

and $\quad \sigma^2 = \text{Var } Z = \text{Var}(aX + bY) = a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2$

$$\therefore \quad M_Z(t) = \exp[t\mu + t^2\sigma^2/2]$$

$$= \exp[t(a\mu_x + b\mu_y) + \frac{t^2}{2}(a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2)]$$

$$= \exp[t_1\mu_x + t_2\mu_y + \frac{1}{2}(t_1^2\sigma_x^2 + 2\rho t_1 t_2\sigma_x\sigma_y + t_2^2\sigma_y^2)] \qquad \ldots(***)$$

where $\quad t_1 = at$ and $t_2 = bt$.

But $(***)$ is the m.g.f. of BVN distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. Hence by uniqueness theorem of m.g.f.

$$(X, Y) \sim BVN\,(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$$

10·10·2. Marginal Distribution of Bivariate Normal Distribution. The marginal distribution of random variable X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy$$

Put $\dfrac{y - \mu_2}{\sigma_2} = u$, then $dy = \sigma_2\, du$. Therefore,

$$f_X(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}}$$

$$\times \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho u\left(\frac{x-\mu_1}{\sigma_1}\right) + u^2\right\}\right]\sigma_2\, du$$

$$= \frac{1}{2\pi\sigma_1\sqrt{(1-\rho^2)}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]$$

$$\times \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{u - \rho\left(\frac{x-\mu_1}{\sigma_1}\right)\right\}^2\right] du$$

Put $\dfrac{1}{\sqrt{(1-\rho^2)}}\left[u - \rho\left(\dfrac{x-\mu_1}{\sigma_1}\right)\right] = t$, then $du = \sqrt{(1-\rho^2)}\; dt$

$$\therefore \quad f_X(x) = \frac{2}{2\pi\sigma_1} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt$$

$$= \frac{1}{2\pi\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \sqrt{2\pi}$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \quad \ldots(10·27)$$

Similarly, we shall get

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx$$

$$= \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right] \quad (10·27a)$$

Hence $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ ...(10·27b)

Remark. We have proved that if $(X, Y) \sim BVN\,(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then the marginal p.d.f.'s of X and Y are also normal. However, the converse is not true, *i.e.,* we may have joint p.d.f. $f(X, Y)$ of (X, Y) which is not normal but the marginal p.d.f.'s may still be normal as discussed in the following illustration.

Consider the joint distribution of X and Y given by :

$$f(x, y) = \frac{1}{2}\left[\frac{1}{2\pi\,(1-\rho^2)^{1/2}} \exp\left\{\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\} + \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 + 2\rho xy + y^2)\right\}\right]$$

$$= \frac{1}{2}[f_1(x, y) + f_2(x, y)]\,;\ -\infty\ (x, y) < \infty \qquad \ldots(10{\cdot}27c)$$

where $f_1(x, y)$ is the p.d.f. of $BVN\ (0, 0, 1, 1, \rho)$ distribution and $f_2(x, y)$ is the p.d.f. of $BVN\ (0, 0, 1, 1, -\rho)$ distribution.

It can be easily verified that $f(x, y)$ is the joint p.d.f. of (X, Y) and obviously $f(x, y)$ is not the p.d.f. of bivariate normal distribution.

Marginal distribution of X in (10·27*c*)

$$f_X(x) = \frac{1}{2}\left[\int_{-\infty}^{\infty} f_1(x, y)\, dy + \int_{-\infty}^{\infty} f_2(x, y)\, dy\right]$$

But $\int_{-\infty}^{\infty} f_1(x, y)\, dy$ is the marginal p.d.f. of X, where

$(X, Y) \sim BVN\ (0, 0, 1, 1, \rho)$ and is given by $X \sim N(0, 1)$.

Similarly $\int_{-\infty}^{\infty} f_2\,(x, y)\, dy$ is the marginal p.d.f. of X, where

$(X, Y) \sim BVN\ (0, 0, 1, 1, -\rho)$ and is given by $X \sim N\ (0, 1)$.

$$\therefore \quad f_X(x) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right]$$

$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,;\ -\infty < x < \infty \qquad \ldots(i)$$

$\Rightarrow X \sim N(0, 1)$ *i.e.*, the marginal distribution of X in (10·27*c*) is normal.

Similarly, we can show that that the marginal p.d.f. of Y in (10·27*c*) is given by :

$$f_Y(y) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-y^2/2} + \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\right]$$

$$= \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,;\ -\infty < y < \infty \qquad \ldots(ii)$$

$\Rightarrow \qquad Y \sim N(0, 1).$

Hence *if the marginal distributions of X and Y are normal (Gaussian), it does not necessarily imply that the joint distribution of (X, Y) is bivariate normal.*

For another illustration, see Question Number 17, Exercise 10(*f*).

We further note that for the joint p.d.f. (10·27*c*), on using (*i*) and (*ii*), we have

$$E(X) = 0,\ \ \sigma_X^2 = 1 \ \text{ and } \ E(Y) = 0,\ \ \sigma_Y^2 = 1.$$

$$\therefore \quad \text{Cov}(X, Y) = \frac{E(XY) - E(X)\,E(Y)}{\sigma_X\,\sigma_Y} = E(XY)$$

$$= \frac{1}{2}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,f_1(x, y)\,dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,f_2(x, y)\,dxdy\right]$$

$$= \frac{1}{2}[\rho + (-\rho)] = 0,$$

because, for $f_1(x, y)$, $(X, Y) \sim BVN\,(0, 0, 1, 1, \rho)$ and for $f_2(x, y)$, $(X, Y) \sim BVN\,(0, 0, 1, 1, -\rho)$.

$$\therefore \quad \text{Corr.}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = 0$$

However, we have ; [From (*i*) and (*ii*)]

$$f_X(x)\,.\,f_Y(y) = \frac{1}{2\pi}\,e^{-\frac{1}{2}(x^2 + y^2)} \neq f(x, y)$$

$\Rightarrow$ X and Y are not independent.

The above example illustrates that we may have a joint density (non-Gaussian) of *rv*'s (X, Y) in which the marginal p.d.f.'s of X and Y are normal and $\rho(X, Y) = 0$ and yet X and Y are not independent.

10·10·3. Conditional Distributions. Conditional distribution of X for a fixed Y is given by

$$f_{X\mid Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_1\sqrt{(1-\rho^2)}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\left(1 - \overline{1-\rho^2}\right)\right\}\right]$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}\sqrt{(1-\rho^2)}} \times \exp\left[-\frac{1}{2(1-\rho^2)\sigma_1^2}\left\{(x-\mu_1) - \rho\,\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right\}^2\right]$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_1\sqrt{(1-\rho^2)}} \times \exp\left[-\frac{1}{2(1-\rho^2)\sigma_1^2}\left\{x - \left(\mu_1 + \rho\,\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right)\right\}^2\right]$$

which is the probability function of a unvariate normal distribution with mean and variance given by

$$E(X \mid Y = y) = \mu_1 + \rho\,\frac{\sigma_1}{\sigma_2}(y - \mu_2) \quad \text{and} \quad V(X \mid Y = y) = \sigma_1^2\,(1 - \rho^2)$$

Hence the conditional distribution of X for fixed Y is given by ;

$$(X \mid Y = y) \sim N\left[\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2 (1 - \rho^2)\right] \quad ...(10\cdot27d)$$

Similarly the conditional distribution of random variables Y for a fixed X is

$$f_{Y \mid X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_2\sqrt{(1 - \rho^2)}}$$

$$\times \exp\left[-\frac{1}{2(1 - \rho^2)\sigma_2^2}\left\{(y - \mu_2) - \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\right\}^2\right],$$

$$-\infty < y < \infty$$

Thus the conditional distribution of Y for fixed X is given by

$$(Y \mid X = x) \sim N\left[\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\ ,\ \sigma_2^2(1 - \rho^2)\right] \quad ...(10\cdot27e)$$

It is apparent from the above results that the array means are collinear, *i.e.*, the regression equations are linear (involving linear functions of the independent variables) and the array variances are constant (*i.e.*, free from independent variable). We express this by saying that *the regression equations of Y on X and X on Y are linear and homoscedastic.*

For $\rho = 0$, the conditional variance $V(Y \mid X)$ is equal to the marginal variance σ_2^2 and the conditional mean $E(Y \mid X)$ is equal to the marginal mean μ_2 and the two variables become independent, which is also apparent from joint distribution function. In between the two extremes when $\rho = \pm 1$, the correlation coefficient ρ provides a measure of degree of association or interdependence between the two variables.

Example 10·27. *Show that for the bivariate normal distribution :*

$$dP = const.\ \exp\left[-\frac{1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2)\right] dx\, dy,$$

(i) M.G.F. is $M(t_1, t_2) = \exp\left[\frac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right]$

(ii) Moments obey the recurrence relation,

$$\mu_{rs} = (r + s - 1)\,\rho\mu_{r-1, s-1} + (r - 1)(s - 1)(1 - \rho^2)\,\mu_{r-2, s-2}$$

Hence or otherwise, show that

$$\mu_{rs} = 0, \text{ if } r + s \text{ is odd}, \mu_{31} = 3\rho, \mu_{22} = 1 + 2\rho^2$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

Solution. (*i*) From the given probability function, we see that

$$\mu_1 = 0 = \mu_2 \quad \text{and} \quad \sigma_1^2 = 1 = \sigma_2^2 .$$

$\therefore$ From (10·26*a*), we get

$$M = M(t_1, t_2) = \exp\left[\tfrac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right]$$

(*ii*) $\quad \dfrac{\partial M}{\partial t_1} = M(t_1 + \rho t_2)$ and $\dfrac{\partial M}{\partial t_2} = M(t_2 + \rho t_1)$

and $$\frac{\partial^2 M}{\partial t_1 \partial t_2} = \frac{\partial}{\partial t_1}\left(\frac{\partial M}{\partial t_2}\right) = \frac{\partial}{\partial t_1}[M(t_2 + \rho t_1)]$$
$$= M\rho + (t_2 + \rho t_1)(t_1 + \rho t_2)M$$

$$\therefore \quad \frac{\partial^2 M}{\partial t_1 \partial t_2} - \rho t_1 \frac{\partial M}{\rho t_1} - \rho t_2 \frac{\partial M}{\partial t_2}$$
$$= [M\rho + (t_2 + \rho t_1)(t_1 + \rho t_2)M - \rho t_1 (t_1 + \rho t_2)M - \rho t_2(t_2 + \rho t_1)M]$$
$$= M[t_1 t_2 + \rho - \rho^2 t_1 t_2] \qquad \textit{(On simplification)}$$
$$= M\rho + (1 - \rho^2) M t_1 t_2$$

$$\therefore \quad \frac{\partial^2 M}{\partial t_1 \partial t_2} = \rho t_1 \frac{\partial M}{\partial t_1} + \rho t_2 \frac{\partial M}{\partial t_2} + M\rho + M\,(1 - \rho^2) t_1 t_2 \qquad \ldots(*)$$

But $$M = \exp\left[\tfrac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right] = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^r\, t_2^s}{r!\, s!}$$

$\therefore$ (*) gives

$$\sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \mu_{rs} \cdot \frac{t_1^{r-1}\, t_2^{s-1}}{(r-1)!\,(s-1)!}$$

$$= \left[\rho \sum_{r=1}^{\infty}\sum_{s=0}^{\infty} r\mu_{rs} \cdot \frac{t_1^r\, t_2^s}{r!\, s!} + \rho \sum_{r=0}^{\infty}\sum_{s=1}^{\infty} s\mu_{rs} \cdot \frac{t_1^r\, t_2^s}{r!\, s!}\right.$$
$$\left. + \rho \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^r\, t_2^s}{r!\, s!} + (1 - \rho^2) \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^{r+1}\, t_2^{s+1}}{r!\, s!}\right]$$

Equating the coefficients of $\dfrac{t_1^{r-1}}{(r-1)!} \cdot \dfrac{t_2^{s-1}}{(s-1)!}$ on both sides, we get

$$\mu_{rs} = [\rho(r-1)\,\mu_{r-1,\,s-1} + \rho(s-1)\mu_{r-1,\,s-1} + \rho^2\mu_{r-1,\,s-1}$$
$$+ (1 - \rho^2)(r-1)(s-1)\mu_{r-2,\,s-2}]$$

$$\Rightarrow \quad \mu_{rs} = (r + s - 1)\,\rho\mu_{r-1,\,s-1} + (r-1)(s-1)(1 - \rho^2)\mu_{r-2,\,s-2}$$

In particular

$$\mu_{31} = 3\rho\mu_{2,0} + 0 = 3\rho\sigma_1^2 = 3\rho \qquad (\because \sigma_1^2 = 1)$$
$$\mu_{22} = 3\rho\mu_{1,1} + (1 - \rho^2)\,\mu_{0,0} = 3\rho^2 + (1 - \rho^2).1$$
$$= (1 + 2\rho^2) \qquad (\because \mu_{11} = \rho\sigma_1\sigma_2 = \rho)$$

Also $$\mu_{03} = \mu_{30} = 0$$
$$\mu_{12} = 2\rho\mu_{0,1} + 0 = 0 \qquad (\because \mu_{01} = \mu_{10} = 0)$$
$$\mu_{23} = 4\rho\mu_{1,2} + 1\cdot 2\,(1 - \rho^2)\mu_{0,1} = 0$$

Similarly, we will get $\mu_{21} = 0$, $\mu_{32} = 0$

If $r + s$ is odd, so is $(r-1) + (s-1)$, $(r-2) + (s-2)$, and so on.

And since $\mu_{30} = 0 = \mu_{03}$, $\mu_{12} = 0 = \mu_{21}$, $\mu_{23} = 0 = \mu_{32} \ldots$, we finally get,

$$\mu_{rs} = 0, \text{ if } r + s \text{ is odd.}$$

Example 10·28. *Show that if X_1 and X_2 are standard normal variates with correlation coefficient ρ between them, then the correlation coefficient between X_1^2 and X_2^2 is given by ρ^2.*

Solution. Since X_1 and X_2, are two standard normal variates, we have

$$E(X_1) = E(X_2) = 0 \text{ and } V(X_1) = E(X_1^2) = 1 = V(X_2) = E(X_2^2)$$

$$\therefore \qquad M_{X_1, X_2}(t_1, t_2) = \exp\left[\tfrac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right] \qquad \text{[c.f. (10·26)]}$$

$$\text{Now} \quad \rho(X_1^2, X_2^2) = \frac{E(X_1^2 X_2^2) - E(X_1^2)\,E(X_2^2)}{\sqrt{[E(X_1^4) - \{E(X_1^2)\}^2]}\ \sqrt{[E(X_2^4) - \{E(X_2^2)\}^2]}}$$

$$\text{where} \quad E(X_1^2 X_2^2) = \text{Coefficient of } \frac{t_1^2}{2!}\cdot\frac{t_2^2}{2!} \text{ in } M(t_1, t_2) = (2\rho^2 + 1)$$

$$E(X_1^4) = \text{Coefficient of } \frac{t_1^4}{4!} \text{ in } M(t_1, t_2) = 3$$

$$E(X_2^4) = \text{Coefficient of } \frac{t_2^4}{4!} \text{ in } M(t_1, t_2) = 3$$

$$\therefore \quad \rho(X_1^2, X_2^2) = \frac{2\rho^2 + 1 - 1}{\sqrt{(3-1)}\sqrt{(3-1)}} = \rho^2$$

Example 10·29. *The variables X and Y with zero means and standard deviations σ_1 and σ_2 are normally correlated with correlation coefficient ρ. Show that U and V defined as*

$$U = \frac{X}{\sigma_1} + \frac{Y}{\sigma_2} \quad \text{and} \quad V = \frac{X}{\sigma_1} - \frac{Y}{\sigma_2}$$

are independent normal variates with variances $2(1 + \rho)$ and $2(1 - \rho)$ respectively.

Solution. We are given that

$$dF(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}\right] dx\,dy$$

$$-\infty < (x, y) < \infty$$

$$\text{Also} \qquad u = \frac{x}{\sigma_1} + \frac{y}{\sigma_2}, \quad v = \frac{x}{\sigma_1} - \frac{y}{\sigma_2}$$

$$\therefore \ \tfrac{1}{2}(u+v) = \frac{x}{\sigma_1} \text{ and } \tfrac{1}{2}(u-v) = \frac{y}{\sigma_2} \Rightarrow x = \frac{\sigma_1}{2}(u+v) \text{ and } y = \frac{\sigma_2}{2}(u-v)$$

Jacobian of transformation J is given by

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{1}{2}\sigma_1 & \frac{1}{2}\sigma_1 \\ \frac{1}{2}\sigma_2 & -\frac{1}{2}\sigma_2 \end{vmatrix} = -\frac{\sigma_1\sigma_2}{2}$$

$$\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} = \frac{1}{4}\left[(u+v)^2 + (u-v)^2\right] = \frac{1}{2}(u^2 + v^2)$$

$$dF_1(u, v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}}$$

$$\times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{1}{2}(u^2 + v^2) - 2\rho\left(\frac{u^2 - v^2}{4}\right)\right\}\right]\frac{\sigma_1\sigma_2}{2}\,du\,dv$$

$$= \frac{1}{2\pi .\ 2\sqrt{(1-\rho^2)}} \exp\left[-\frac{1}{4(1-\rho^2)}\{(1-\rho)u^2+(1+\rho)v^2\}\right] du\, dv$$

$$= \frac{1}{2\pi\sqrt{2(1-\rho)}\sqrt{2(1+\rho)}} \cdot \exp\left[-\frac{u^2}{2(1+\rho)2} - \frac{v^2}{2(1-\rho)2}\right] du\, dv$$

$$= \left[\frac{1}{\sqrt{2\pi}\sqrt{2(1+\rho)}} \cdot \exp\left\{-\frac{u^2}{2(1+\rho)2}\right\}\right] du$$

$$\times \left[\frac{1}{\sqrt{2\pi}\sqrt{2(1-\rho)}} \cdot \exp\left\{-\frac{v^2}{2(1-\rho)2}\right\}\right] dv$$

$= [f_1(u)du]\ [f_2(v)dv]$, (say)

where $$f_1(u) = \frac{1}{\sqrt{2\pi}\sqrt{2(1+\rho)}} .\exp\left\{-\frac{u^2}{2(1+\rho)2}\right\}$$

and $$f_2(v) = \frac{1}{\sqrt{2\pi}\sqrt{2(1-\rho)}} .\exp\left\{-\frac{v^2}{2(1-\rho)2}\right\}$$

Hence U and V are independently distributed, U as $N\,[0, 2(1+\rho)]$ and V as $N\,[0, 2\,(1-\rho)]$.

Aliter. Find joint m.g.f. of U and V viz.,

$$M\,(t_1, t_2) = E\,(e^{t_1U + t_2V}) = E\,\left[e^{X(t_1+t_2)/\sigma_1 + Y(t_1-t_2)/\sigma_2}\right]$$

and use $E(e^{t_1X + t_2Y}) = \exp\,[(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t_1 t_2\sigma_1\sigma_2)/2]$

Example 10·30. *If X and Y are standard normal variates with co-efficient of correlation* ρ, *show that*

(*i*) *Regression of Y on X is linear.*

(*ii*) $X + Y$ and $X - Y$ *are independently distributed.*

(*iii*) $Q = \dfrac{X^2 - 2\rho XY + Y^2}{(1-\rho^2)}$ *is distributed like a chi-square, i.e., as that of the sum of the squares of standard normal variates.*

(*Madras Univ. B.E., 1990*)

Solution. (*i*) c.f. § 10·10·3.

(*ii*) Let $u = x + y$ and $v = x - y$

$$dF\,(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right] dxdy$$

Now $$x = \frac{u+v}{2},\ y = \frac{u-v}{2}$$

$$\therefore \quad J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{2} & \dfrac{1}{2} \\ \dfrac{1}{2} & -\dfrac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

$$dG(u, v) = C \exp\left[-\frac{1}{2(1-\rho^2)\,.\,4}\{2(u^2+v^2) - 2\rho(u^2-v^2)\}\right] dudv$$

where $\quad C = \dfrac{1}{4\pi\sqrt{1-\rho^2}}$

$$\therefore dG(u, v) = C \exp\left[-\frac{1}{4(1-\rho^2)}\{(1-\rho)u^2 + (1+\rho)v^2\}\right]dudv$$

$$= \left[C_1 \exp\left\{-\frac{u^2}{4(1+\rho)}\right\}du\right] \times \left[C_2 \exp\left\{-\frac{v^2}{4(1-\rho)}\right\}dv\right]$$

$$= [g_1(u)du]\,[g_2(v)dv], \text{ (say)}.$$

Hence U and V are independently distributed.

(iii) $$M_Q(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{tQ}\, dF(x, y)$$

$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(tQ)$$

$$\times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{x^2 - 2\rho xy + y^2\right\}\right]dxdy$$

$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(tQ - \frac{Q}{2}\right)dxdy$$

$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\frac{Q}{2}(1-2t)\right]dxdy$$

Put $\sqrt{(1-2t)}\,x = u$ and $\sqrt{(1-2t)}\,y = v$

$\therefore \quad dx = \dfrac{du}{\sqrt{(1-2t)}}$ and $dy = \dfrac{dv}{\sqrt{(1-2t)}}$

Also $\quad Q = \dfrac{1}{(1-\rho^2)}[x^2 - 2\rho xy + y^2] = \dfrac{1}{(1-\rho^2)}\left[\dfrac{u^2 - 2\rho uv + v^2}{1-2t}\right]$

$$\therefore \quad M_Q(t) = \frac{1}{2\pi\sqrt{(1-\rho^2)}\,(1-2t)}$$

$$\times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)\right]du\,dv$$

$$= \frac{1}{(1-2t)}\cdot 1 = (1-2t)^{-1}$$

which is the m.g.f. of chi-square (χ^2) variate* with n (=2) degrees of freedom.

Example 10·31. *Let X and Y be independent standard normal variates. Obtain the m.g.f. of XY.* [*Gauhati Univ. M.Sc., 1992*]

Solution. We have, by definition :

$$M_{XY}(t) = E(e^{tXY}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{txy} \cdot f(x, y)\, dxdy$$

Since X and Y are independent standard normal variates, their joint p.d.f. $f(x, y)$ is given by :

$$f(x, y) = f_1(x) \cdot f_2(y) = \frac{1}{2\pi} e^{-x^2/2}\, e^{-y^2/2}\ ;\ -\infty < (x, y) < \infty$$

$$\therefore\quad M_{XY}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2 - 2txy + y^2)}\, dxdy$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\Big[-\frac{1}{2(1-t^2)}$$

$$\times \Big\{\frac{x^2}{1/(1-t^2)} - \frac{2t.x.y}{(1/\sqrt{1-t^2})(1/\sqrt{1-t^2})} + \frac{y^2}{1/(1-t^2)}\Big\}\Big]\, dxdy \qquad \ldots(*)$$

If $(U, V) \sim BVN\,(0, 0, \sigma_1^2, \sigma_2^2, \rho)$, then we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}}\, dxdy = 1$$

$$\Rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}}\, dxdy = 2\pi\sigma_1\sigma_2\sqrt{1-\rho^2} \qquad \ldots(**)$$

Comparing (*) and (**) with

$$\sigma_1^2 = \sigma_2^2 = \frac{1}{(1-t^2)} \quad \text{and} \quad \rho = t, \text{ we get}$$

$$M_{XY}(t) = \frac{1}{2\pi} \cdot 2\pi \frac{1}{\sqrt{1-t^2}} \cdot \frac{1}{\sqrt{1-t^2}} \cdot \sqrt{1-t^2}$$

$$\Rightarrow \quad M_{XY}(t) = (1-t^2)^{1/2}\ ;\ -1 < t < 1$$

Example 10·32. *Let X and Y have bivariate normal distribution with parameters :*

$\mu_X = 5, \mu_Y = 10, \sigma_X^2 = 1, \sigma_Y^2 = 25$ and Corr (X, Y) = ρ.

(a) If $\rho > 0$, find ρ when $P\,(4 < Y < 16 \mid X = 5) = 0{\cdot}954$

[*Delhi Univ. B.Sc. (Math. Hons.), 1993, '83*]

---

*Chi-square distribution is discussed in Chapter 13 *(b) If $\rho = 0$, find $P(X + Y \le 16)$.*

Solution. Since $(X, Y) \sim BVN\ (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, the conditional distribution of Y given $X = x$ is also normal.

$$(Y \mid X = x) \sim N\left[\mu = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X),\ \sigma^2 = \sigma_Y^2(1 - \rho^2)\right]$$

$$\therefore\quad (Y \mid X = 5) \sim N\left[\mu = 10 + \rho \times \frac{5}{1}(5 - 5),\ \sigma^2 = 25(1 - \rho^2)\right]$$

$$= N\left[\mu = 10,\ \sigma^2 = 25(1 - \rho^2)\right]$$

We want ρ so that

$$P(4 < Y < 16 \mid X = 5) = 0{\cdot}954$$

where
$$Z = \frac{Y - \mu}{\sigma} = \frac{Y - 10}{5\sqrt{(1 - \rho^2)}} \sim N(0, 1)$$

$$\Rightarrow\quad P\left(\frac{4 - 10}{\sigma} < Z < \frac{16 - 10}{\sigma}\right) = 0{\cdot}954$$

$$\Rightarrow\quad P\left(\frac{-6}{\sigma} < Z < \frac{6}{\sigma}\right) = 0{\cdot}954 \qquad \ldots(*)$$

But we know that if $Z \sim N(0, 1)$, then

$$P(-2 < Z < 2) = 0{\cdot}954 \qquad \ldots(**)$$

Comparing (*) and (**), we get

$$\frac{6}{\sigma} = 2 \;\Rightarrow\; \sigma = 3 \;\Rightarrow\; \sigma^2 = 9 = 25(1 - \rho^2)$$

$$\therefore\quad 1 - \rho^2 = \frac{9}{25} \;\Rightarrow\; \rho^2 = \frac{16}{25} \;\Rightarrow\; \rho = \frac{4}{5} = 0{\cdot}8 \qquad (\because \rho > 0)$$

(*b*) Since (X, Y) have bivariate normal distribution,

$$\rho = 0 \Rightarrow X \text{ and } Y \text{ are independent } rv\text{'s}$$

and
$$X \sim N(\mu_X, \sigma_X^2) \text{ and } Y \sim N(\mu_Y, \sigma_Y^2)$$

$$\therefore\quad X + Y \sim N(\mu = \mu_X + \mu_Y,\ \sigma^2 = \sigma_X^2 + \sigma_Y^2) = N(15, 26)$$

Hence

$$P(X + Y \le 16) = P\left(Z \le \frac{16 - 15}{\sqrt{26}}\right)$$

where $Z = \dfrac{(X + Y) - \mu}{\sigma} \sim N(0, 1)$.

$$\therefore\quad P(X + Y \le 16) = P\left(Z \le \frac{1}{\sqrt{26}}\right) = \Phi\left(1/\sqrt{26}\right),$$

where $\Phi(z) = P(Z \le z)$, is the distribution function of standard normal variate.

Remark.
$$P(X + Y \le 16) = P\left(Z \le \frac{1}{5{\cdot}099}\right) = P(Z \le 0{\cdot}196)$$
$$= 0{\cdot}5 + P(0 \le Z \le 0{\cdot}196)$$
$$= 0{\cdot}5 + 0{\cdot}0793 \text{ (approx.)}$$
$$= 0{\cdot}5793.$$

EXERCISE 10(f)

1. (*a*) Define conditional and marginal distributions. If X and Y follow bivariate normal distribution, find (*i*) the conditional distribution of X given Y and (*ii*) the marginal distribution of X. Show that the conditional mean of X is dependent on the given Y, but the conditional variance is independent of it.

(*b*) Define Bivariate Normal distribution. If (X, Y) has a bivariate normal distribution, find the marginal density function $f_X(x)$ of X.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1988*]

2. (*a*) The marks X and Y scored by candidates in an examination in two subjects Mathematics and Statistics are known to follow a bivariate normal distribution. The mean of X is 52 and its standard deviation is 15, while Y has mean 48 and standard deviation 13. Also the coefficient of correlation between X and Y is 0·6.

Write down the joint distribution of X and Y. If 100 marks in the aggregate are needed for a pass in the examination, show how to calculate the proportion of candidates who pass the examination ?

(*b*) A manufacturer of electric bulbs, in his desire for putting only good bulbs for sale, rejects all bulbs for which a certain quality characteristic X of the filament is less than 65 units. Assume that the quality characteristic X and the life Y, of the bulb in hours are jointly normally distributed with parameters given below :

	X	Y
Mean	80	1100
Standard deviation	10	10

Correlation coefficient $\rho(X, Y) = 0{\cdot}60$

Find (*i*) the proportion of bulbs produced that will burn for less than 1000 hours, (*ii*) the proportion of bulbs produced that will be put for sale, (*iii*) the average life of bulbs put for sale.

3. (*a*) Determine the parameters of the bivariate normal distribution :

$$f(x, y) = k \exp\left[-\frac{8}{27}\{(x-7)^2 - 2(x-7)(y+5) + 4(y+5)^2\}\right]$$

Also find the value of k.

(*b*) For the bivariate normal distribution :

$$(X, Y) \sim BVN\left(1, 2, 4^2, 5^2, \tfrac{12}{13}\right)$$

find (*i*) $P(X > 2)$, (*ii*) $P(X > 2 \mid Y = 2)$.

(*c*) The bivariate random variable (X_1, X_2) have a bivariate normal distribution with means 60 and 75 and standard deviations 6 and 12 with a correlation coefficient of 0·55. Find the following probabilities :

(*i*) $P(65 \le X_1 \le 75)$, (*ii*) $P(71 \le X_2 \le 80 \mid X_1 = 55)$ and (*iii*) $P(|X_1 - X_2| \ge 15)$.

4. For a bivariate normal distribution :

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\},$$

$$-\infty < (x, y) < \infty$$

Find (i) marginal distribution of X and Y,

(ii) conditional distribution of Y given X,

(iii) distribution of $\dfrac{1}{(1-\rho^2)}[x^2 - 2\rho xy + y^2]$,

and (iv) show that in general X and Y are stochastically dependent and will be independent if and only if $\rho = 0$.

5. Let the joint p.d.f. of X and Y be

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}}$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

where $-\infty < x < \infty, -\infty < y < \infty, -1 < \rho < 1$.

(i) Find the marginal distribution of X.

(ii) Find the conditional distribution of Y given $X = x$.

(iii) Show that the regression of Y on X is linear and homoscedastic.

(iv) Find $P\{3 < Y < 8 \mid X = 7\}$, given that $\mu_1 = 3$, $\mu_2 = 1$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$, $\rho = 0{\cdot}6$,

(v) Find the probability of the simultaneous materialization of the inequalities, $X > E(X)$ and $Y > E(Y)$

Hint. (v) Required probability p is given by

$$p = P[X > E(X), Y > E(Y)] = P[X > \mu_1) \cap (Y > \mu_2)]$$

$$= \int_{\mu_1}^{\infty}\int_{\mu_2}^{\infty} f(x, y)\, dx\, dy$$

$$= \int_0^{\infty}\int_0^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)\right] du\, dv,$$

$$\left(u = \frac{x-\mu_1}{\sigma_1},\ v = \frac{y-\mu_2}{\sigma_2}\right).$$

Now proceed as in Hint to Question Number 9(b).

6. Let the random variables X and Y be assumed to have a joint bivariate normal distribution with

$$\mu_1 = \mu_2 = 0,\ \sigma_1 = 4,\ \sigma_2 = 3 \text{ and } r(X, Y) = 0{\cdot}8.$$

(i) Write down the joint density function of X and Y.

(ii) Write down the regression of Y on X.

(iii) Obtain the joint density of $X + Y$ and $X - Y$.

7. For the distribution of random variables X and Y given by

$$dF = k \exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right] dx\, dy;\ -\infty \le x \le \infty,\ -\infty \le y \le \infty.$$

Obtain

(i) the constant k,

(ii) the distributions of X and Y,

(iii) the distributions of X for given Y and of Y for given X,

(iv) the curves of regression of Y on X and of X on Y,

and (v) the distributions of $X + Y$ and $X - Y$.

8. Let (X, Y) be a bivariate normal random variable with $E(X) = E(Y) = 0$, Var (X) = Var $(Y) = 1$ and Cov $(X, Y) = \rho$. Show that the random variable $Z = Y/X$ has a Cauchy distribution.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1989*]

Ans. $f(z) = \dfrac{1}{\pi}\left[\dfrac{(1-\rho^2)^{1/2}}{(1-\rho^2)+(z-\rho)^2}\right], -\infty < z < \infty.$

9. (a) If $(X, Y) \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, prove that

$$P(X > \mu_x \cap Y > \mu_y) = \frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}$$

[*Delhi Univ. M.Sc. (Stat.), 1987*]

(b) If $(X, Y) \sim N(0, 0, 1, 1, \rho)$ then prove that

$$P(X > 0 \cap Y > 0) = \frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}.$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

Hint. $p = P(X > 0 \cap Y > 0)$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \times \int_0^\infty \int_0^\infty \exp\left[-\frac{1}{2(1-\rho^2)}\{x^2 - 2\rho xy + y^2\}\right] dx\, dy$$

Put $x = r\cos\theta$, $y = r\sin\theta \Rightarrow |J| = r\,;\, 0 < r < \infty, 0 \le \theta \le \pi/2$

$$\therefore\ p = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_0^\infty \int_0^{\pi/2} \text{xp}\left[-\frac{r^2}{2(1-\rho^2)}(1 - \rho\sin 2\theta)\right] r\, dr\, d\theta$$

Now integrate first w.r. to r and then w.r. to θ.

10. (a) Let X_1 and X_2 be two independent normally distributed variables with zero means and unit variances. Let Y_1 and Y_2 be the linear functions of X_1 and X_2 defined by

$$Y_1 = m_1 + l_{11}X_1 + l_{12}X_2,\ Y_2 = m_2 + l_{21}X_1 + l_{22}X_2$$

Show that Y_1 and Y_2 are normally distributed with means m_1 and m_2, variances $\mu_{20} = l_{11}^2 + l_{12}^2$, $\mu_{02} = l_{21}^2 + l_{22}^2$, and covariance $\mu_{11} = l_{11}l_{21} + l_{12}l_{22}$.

(b) Let X_1 and X_2 be independent standard normal variates. Show that the variates Y_1, Y_2 defined by

$Y_1 = a_1 + b_{11}X_1 + b_{12}X_2$, $Y_2 = a_2 + b_{21}X_1 + b_{22}X_2$ are dependent normal variates and find their mean and variance.

Hint. Y_1 and Y_2, being linear combination of S.N.V's are also normally distributed. To prove that they are dependent, it is sufficient to prove that $r(Y_1, Y_2) \ne 0$. [*c.f.* Remark 2 to Theorem 10·2)

11. (*a*) Show that, if X and Y are independent normal variates with zero means and variances σ_1^2 and σ_2^2 respectively, the point of inflexion of the curve of intersection of the normal correlation surface by planes through the z-axis, lie on the elliptical cylinder,

$$\frac{X^2}{\sigma_1^2} + \frac{Y^2}{\sigma_2^2} = 1$$

(*b*) If X and Y are bivariate normal variates with standard deviations unity and with correlation coefficient ρ, show that the regression of X^2 (Y^2) on Y^2 (X^2) is strictly linear. Also show that the regression of X (Y) on Y^2 (X^2) is not linear.

12. For the bivariate normal distribution :

$$dF = k \exp\left[-\tfrac{2}{3}(x^2 - xy + y^2 - 3x + 3y + 3)\right] dx\, dy,$$

obtain (*i*) the marginal distribution of Y, and

(*ii*) the conditional distribution of Y given X.

Also obtain the characteristic function of the above bivariate normal ditribution and hence the covariance between X and Y.

13. Let f and g be the p.d.f.'s with corresponding distribution functions F and G. Also let

$$h(x, y) = f(x)\, g(y)\, [1 + \alpha\, (2F(x) - 1)\, (2G(y) - 1)],$$

where $|\alpha| \le 1$, is a constant and h is a bivariate p.d.f. with marginal p.d.f.'s f and g. Further let f and g be p.d.f.'s of $N(0, 1)$ distribution. Then prove that :

$$\text{Cov}(X, Y) = \alpha/\pi$$

14. If $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, compute the correlation coefficient between e^X and e^Y.

Hint. Let $U = e^X$, $V = e^Y$.

$$\mu'_{rs} = E(U^r.V^s) = E\left[e^{rX + sY}\right]$$

$$= \exp\left[r\mu_1 + s\mu_2 + \tfrac{1}{2}(r^2\sigma_1^2 + s^2\sigma_2^2 + 2\rho rs)\right]$$

[c.f. m.g.f. of $B.V.N.$ distribution : $t_1 = r$, $t_2 = s$]

Now $E(U) = \mu'_{10}$; $E(U^2) = \mu'_{20}$, $E(UV) = \mu_{11}'$ and so on.

Ans. $$\rho(U,V) = \frac{e^{\rho\sigma_1\sigma_2} - 1}{[(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)]^{1/2}}$$

15. If $(X, Y) \sim BVN(0, 0, 1, 1, \rho)$, find $E[\max(X, Y)]$.

Hint. $\max(X, Y) = \frac{1}{2}(X + Y) + \frac{1}{2}|X - Y|$

and $Z = X - Y \sim N[0, 2(1 - \rho)]$ [c.f.Theorem 10·6]

Ans. $E[\max(X, Y)] = \left(\dfrac{1 - \rho}{\pi}\right)^{1/2}$

16. If $(X, Y) \sim BVN(0, 0, 1, 1, \rho)$ with joint p.d.f. $f(x, y)$ then prove that

(*a*) $P(XY > 0) = \frac{1}{2} + \frac{1}{\pi} \sin^{-1}(\rho)$.

Hint. $P(XY > 0) = P(X > 0 \cap Y > 0) + P(X < 0 \cap Y < 0)$
$= 2\,P(X > 0 \cap Y > 0)$ [By symmetry]

Now proceed as in Hint to Question No. 9(*b*).

(*b*) $$2\pi \int_{-\infty}^{0} \int_{-\infty}^{0} f(x, y)\, dx dy = \pi + \sin^{-1} \rho$$

17. The joint density of *r.v's* (X, Y) is given by :

$$f(x, y) = \frac{1}{2\pi} \cdot \exp\,[-(x^2 + y^2)/2] \times [1 + xy \exp\,\{-(x^2 + y^2 - 2)/2\}]\,;$$
$$-\infty < (x, y) < \infty$$

(*i*) Verify that $f(x, y)$ is a p.d.f.

(*ii*) Show that the marginal distribution of each of X and Y is normal.

(*iii*) Are X and Y independent ?

Ans. (*ii*) $X \sim N(0, 1)$, $Y \sim N(0, 1)$

(*ii*) X and Y are not independent.

18. Show by means of an example that the normality of conditional p.d.f.'s does not imply that the bivariate density is normal.

Hint. Consider $f(x, y) = \text{constant.} \exp\,[-(1 + x^2)(1 + y^2)]$; $-\infty < (x, y) < \infty$

Then $(Y \mid x) \sim N\left(0, \dfrac{1}{2(1 + x^2)}\right)$ and $(X \mid y) \sim N\left(0, \dfrac{1}{2(1 + y^2)}\right)$

19. For a bivariate normal *r.v.* (X, Y), does the conditional p.d.f. of (X, Y) given $X + Y = c$, (constant) exist ? If so find it. If not, why not ?

Ans. No, since $P(X + Y = c) = 0$.

20. Let

$$f(x, y) = \frac{1}{2}\left[\frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2)\right\} + \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)}(x^2 + 2\rho xy + x^2)\right\}\right]$$
$$-\infty < x < \infty, \quad -\infty < y < \infty$$

then show that :

(*i*) $f(x, y)$ is a joint p.d.f. such that both marginal densities are normal but $f(x, y)$ is not bivariate normal.

(*ii*) X and Y have zero correlation but X and Y are not independent.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

21. Let X, Y be normally correlated variates with zero means and variances σ_1^2, σ_2^2 and if

$$W = \frac{X}{\sigma_1}, \; Z = \frac{1}{\sqrt{(1 - \rho^2)}}\left\{\frac{Y}{\sigma_2} - \frac{\rho X}{\sigma_1}\right\}$$

Show that $$\frac{\partial(w, z)}{\partial(x, y)} = \frac{1}{\sigma_1\sigma_2\sqrt{(1 - \rho^2)}}$$

and $$W^2 + Z^2 = \frac{1}{(1-\rho^2)}\left[\frac{X^2}{\sigma_1^2} - \frac{2\rho XY}{\sigma_1\sigma_2} + \frac{Y^2}{\sigma_2^2}\right]$$

Deduce that the joint probability differential of W and Z is

$$dP = \frac{1}{2\pi} \cdot \exp\left[-\frac{1}{2}(w^2+z^2)\right] dw\,dz$$

and hence that W, Z are independent normal variates with zero means and unit S.D.'s [*Meerut Univ. M.Sc., 1993*]

Hence or otherwise obtain the m.g.f. of the bivariate normal distribution.

22. From a standard bivariate normal population, a random sample of n observations (X_i, Y_i), $(i = 1, 2, \ldots, n)$ is drawn. Show that the distribution of

$$Z_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \text{ and } Z_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

has the moment generating function :

$$\text{Constant}\left[\left(1-\frac{2t_1}{n}\right)\left(1-\frac{2t_2}{n}\right) - \frac{4\rho^2 t_1 t_2}{n^2}\right]^{-n/2}$$

Hint. $$M_{Z_1, Z_2}(t_1, t_2) = \left[E\exp\left(\frac{t_1 x^2}{n} + \frac{t_2 y^2}{n}\right)\right]^n$$

$$= \left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[x^2\left(\frac{t_1}{n} - \frac{1}{2(1-\rho^2)}\right) + \left(\frac{\rho}{1-\rho^2}\right)xy + y^2\left(\frac{t_2}{n} - \frac{1}{2(1-\rho^2)}\right)\right] dxdy\right\}^n$$

Now use the result

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp[-(ax^2 + 2hxy + by^2)]\,dxdy = \frac{\pi}{\sqrt{ab-h^2}}$$

and simplify.

10·11. Multiple and Partial Correlation. When the values of one variable are associated with or influenced by other variable, *e.g.*, the age of husband and wife, the height of father and son, the supply and demand of a commodity and so on, Karl Pearson's coefficient of correlation can be used as a measure of linear relationship between them. But sometimes there is interrelation between many variables and the value of one variable may be influenced by many others, *e.g.*, the yield of crop per acre say (X_1) depends upon quality of seed (X_2), fertility of soil (X_3), fetilizer used (X_4), irrigation facilities (X_5), weather conditions (X_6) and so on. Whenever we are interested in studying the joint effect of a group of variables upon a variable not included in that group, our study is that of *multiple correlation and multiple regression.*

Suppose in a trivariate or multi-variate distribution we are interested in the relationship between two variables only. The are two alternatives, *viz.*, (*i*) we consider only those two members of the observed data in which the other members have specified values or (*ii*) we may eliminate mathematically the effect of other variates on two variates. The first method has the disadvantage that it limits the size of the data and also it will be applicable to only the data in which the other variates have assigned values. In the second method it may not be possible to eliminate the entire influence of the variates but the linear effect can be easily eliminated. The correlation and regression between only two variates eliminating the linear effect of other variates in them is called the *partial correlation and partial regression.*

10·11·1. Yule's Notation. Let us consider a distribution involving three random variables X_1, X_2 and X_3. Then the equation of the plane of regression of X_1 on X_2 and X_3 is

$$X_1 = a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3 \qquad \ldots(10\cdot28)$$

Without loss of generality we can assume that the variables X_1, X_2 and X_3 have been measured from their respective means, so that

$$E(X_1) = E(X_2) = E(X_3) = 0$$

Hence on taking expectation of both sides in (10·28), we get $a = 0$.

Thus the plane of regression of X_1 on X_2 and X_3 becomes

$$X_1 = b_{12\cdot3}\,X_2 + b_{13\cdot2}\,X_3 \qquad \ldots(10\cdot28a)$$

The coefficients $b_{12\cdot3}$ and $b_{13\cdot2}$ are known as the *partial regression coefficients* of X_1 on X_2 and of X_1 on X_3 respectively.

$$e_{1\cdot23} = b_{12\cdot3}\,X_2 + b_{13\cdot2}\,X_3$$

is called the estimate of X_1 as given by the plane of regression (10·28*a*) and the quantity

$$X_{1\cdot23} = X_1 - b_{12\cdot3}\,X_2 - b_{13\cdot2}\,X_3,$$

is called the *error of estimate* or *residual.*

In the general case of n variables $X_1, X_2, \ldots, X_n$ the equation of the plane of regression of X_1 on $X_2, X_3, \ldots, X_n$ becomes

$$X_1 = b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\ldots(n-1)}\,X_n$$

The error of estimate or residual is given by

$$X_{1\cdot23\ldots n} = X_1 - (b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\ldots(n-1)}\,X_n)$$

The notations used here are due to Yule. The subscripts before the dot (.) are known as *primary subscripts* and those after the dot are called *secondary subscripts.* The order of a regression coefficient is determined by the number of secondary subscripts, *e.g.*,

$$b_{12\cdot3}, b_{12\cdot34}, \ldots, b_{12\cdot34\ldots n}$$

are the regression coefficients of order 1, 2, ... $(n-2)$ respectively. Thus in general, a regression coefficient with p-secondary subscripts will be called a regression co-efficient of order 'p'. It may be pointed out that the order in which the secondary subscripts are written is immaterial but the order of the primary subscripts is important, *e.g.*, in $b_{12\cdot34\ldots n}$, X_2 is independent while X_1 is dependent variable but in $b_{21\cdot34\ldots n}$, X_1 is independent while X_2 is dependent variable. Thus of the two primary subscripts, former refers to dependent variable and the latter to independent variable.

The order of a residual is also determined by the number of secondary subscripts in it, *e.g.*, $X_{1.23}, X_{1.234}, \ldots, X_{1.23\ldots n}$ are the residuals of order 2, 3, ..., $(n-1)$ respectively.

Remark. In the following sequences we shall assume that the variables under consideration have been measured from their respective means.

10·12. Plane of Regression. The equation of the plane of regression of X_1 on X_2 and X_3 is

$$X_1 = b_{12.3} X_2 + b_{13.2} X_3 \qquad \ldots(10\cdot29)$$

The constants b's in (10·29) are determined by the principle of least squares, *i.e.*, by minimising the sum of the squares of the residuals, *viz.*,

$$S = \Sigma X_{1.23}^2 = \Sigma(X_1 - b_{12.3} X_2 - b_{13.2} X_3)^2,$$

the summation being extended to the given values (N in number) of the variables.

The normal equations for estimating $b_{12.3}$ and $b_{13.2}$ are

$$\left.\begin{aligned} \frac{\partial S}{\partial b_{12.3}} &= 0 = -2\,\Sigma X_2(X_1 - b_{12.3} X_2 - b_{13.2} X_3) \\ \frac{\partial S}{\partial b_{13.2}} &= 0 = -2\,\Sigma X_3(X_1 - b_{12.3} X_2 - b_{13.2} X_3) \end{aligned}\right\} \qquad \ldots(10\cdot30)$$

i.e., $$\Sigma X_2 X_{1.23} = 0 \quad \text{and} \quad \Sigma X_3 X_{1.23} = 0 \qquad \ldots(10\cdot30a)$$

$$\Rightarrow \quad \left.\begin{aligned} \Sigma X_1X_2 - b_{12.3}\,\Sigma X_2^2 - b_{13.2}\,\Sigma X_2X_3 &= 0 \\ \Sigma X_1X_3 - b_{12.3}\,\Sigma X_2X_3 - b_{13.2}\,\Sigma X_3^2 &= 0 \end{aligned}\right\} \qquad \ldots(10\cdot30b)$$

Since X_i's are measured from their respective means, we have

$$\left.\begin{aligned} \sigma_i^2 &= \frac{1}{N}\Sigma X_i^2, \ \text{Cov}\,(X_i, X_j) = \frac{1}{N}\Sigma X_i X_j \\ \text{and} \qquad r_{ij} &= \frac{\text{Cov}\,(X_i, X_j)}{\sigma_i\sigma_j} = \frac{\Sigma X_i X_j}{N\,\sigma_i\sigma_j} \end{aligned}\right\} \qquad \ldots(10\cdot30c)$$

Hence from (10·30*b*), we get

$$\left.\begin{aligned} r_{12}\,\sigma_1\sigma_2 - b_{12.3}\,\sigma_2^2 - b_{13.2}\,r_{23}\,\sigma_2\sigma_3 &= 0 \\ r_{13}\,\sigma_1\sigma_3 - b_{12.3}\,r_{23}\sigma_2\sigma_3 - b_{13.2}\,\sigma_3^2 &= 0 \end{aligned}\right\} \qquad \ldots(10\cdot30d)$$

Solving the equations (10·30*d*) for $b_{12.3}$ and $b_{13.2}$, we get

$$b_{12.3} = \frac{\begin{vmatrix} r_{12}\sigma_1 & r_{23}\,\sigma_3 \\ r_{13}\,\sigma_1 & \sigma_3 \end{vmatrix}}{\begin{vmatrix} \sigma_2 & r_{23}\,\sigma_3 \\ r_{23}\,\sigma_2 & \sigma_3 \end{vmatrix}} = \frac{\sigma_1}{\sigma_2}\cdot\frac{\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} \qquad \ldots(10\cdot31)$$

Similarly, we will get

$$b_{13\cdot2}=\frac{\sigma_1}{\sigma_3}\cdot\frac{\begin{vmatrix}1 & r_{12}\\ r_{23} & r_{13}\end{vmatrix}}{\begin{vmatrix}1 & r_{23}\\ r_{23} & 1\end{vmatrix}} \qquad \text{...(10·31}a\text{)}$$

If we write

$$\omega=\begin{vmatrix}1 & r_{12} & r_{13}\\ r_{21} & 1 & r_{23}\\ r_{31} & r_{32} & 1\end{vmatrix} \qquad \text{...(10·32)}$$

and ω_{ij} is the cofactor of the element in the ith row and jth column of ω, we have from (10·31) and (10·31a)

$$b_{12\cdot3}=-\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}} \quad\text{and}\quad b_{13\cdot2}=-\frac{\sigma_1}{\sigma_3}\cdot\frac{\omega_{13}}{\omega_{11}} \qquad \text{...(10·33)}$$

Substituting these values in (10·29), we get the required equation of the plane of regression of X_1 on X_2 and X_3 as

$$X_1=-\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}}\cdot X_2-\frac{\sigma_1}{\sigma_3}\cdot\frac{\omega_{13}}{\omega_{11}}\cdot X_3$$

$$\Rightarrow \qquad \frac{X_1}{\sigma_1}\cdot\omega_{11}+\frac{X_2}{\sigma_2}\cdot\omega_{12}+\frac{X_3}{\sigma_3}\cdot\omega_{13}=0 \qquad \text{...(10·34)}$$

Aliter. Eliminating the coefficient $b_{12\cdot3}$ and $b_{13\cdot2}$ in (10·29) and (10·30d), the required equation of the plane of regression of X_1 on X_2 and X_3 becomes

$$\begin{vmatrix}X_1 & X_2 & X_3\\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & r_{23}\sigma_2\sigma_3\\ r_{13}\sigma_1\sigma_3 & r_{23}\sigma_2\sigma_3 & \sigma_3^2\end{vmatrix}=0$$

Dividing C_1, C_2 and C_3 by σ_1, σ_2 and σ_3 respectively and also R_2 and R_3 by σ_2 and σ_3 respectively, we get

$$\begin{vmatrix}\dfrac{X_1}{\sigma_1} & \dfrac{X_2}{\sigma_2} & \dfrac{X_3}{\sigma_3}\\ r_{12} & 1 & r_{23}\\ r_{13} & r_{23} & 1\end{vmatrix}=0$$

$$\Rightarrow \qquad \frac{X_1}{\sigma_1}\omega_{11}+\frac{X_2}{\sigma_2}\omega_{12}+\frac{X_3}{\sigma_3}\omega_{13}=0$$

where ω_{ij} is defined in (10·32).

10·12·1. Generalisation. In general, the equation of the plane of regression of X_1 on $X_2, X_3, \ldots X_n$ is

$$X_1=b_{12\cdot34\ldots n}X_2+b_{13\cdot24\ldots n}X_3+\ldots+b_{1n\cdot23\ldots(n-1)}X_n \qquad \text{...(10·35)}$$

The sum of the squares of residuals is given by

$$S=\Sigma X^2_{1\cdot23\ldots n}$$

$$= \sum (X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n)^2$$

Using the principle of least squares, the normal equations for estimating the $(n-1)$, b's are

$$\left.\begin{aligned}
\frac{\partial S}{\partial b_{12\cdot34\ldots n}} &= 0 = -2\sum X_2(X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n)\\
\frac{\partial S}{\partial b_{13\cdot24\ldots n}} &= 0 = -2\sum X_3(X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n)\\
&\vdots\\
\frac{\partial S}{\partial b_{1n\cdot23\ldots(n-1)}} &= 0 = -2\sum X_n(X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n)
\end{aligned}\right]$$

...(10·36)

i.e., $\sum X_i X_{1\cdot23\ldots n} = 0, \ (i = 2, 3, \ldots, n)$ ...(10·36*a*)

which on smplification after using (10·30*c*), give

$$\left.\begin{aligned}
r_{12}\sigma_1\sigma_2 &= b_{12\cdot34\ldots n}\sigma_2^2 + b_{13\cdot24\ldots n} r_{23}\sigma_2\sigma_3 + \ldots + b_{1n\cdot23\ldots(n-1)} r_{2n}\sigma_2\sigma_n\\
r_{13}\sigma_1\sigma_3 &= b_{12\cdot34\ldots n} r_{23}\sigma_2\sigma_3 + b_{13\cdot24\ldots n}\sigma_3^2 + \ldots + b_{1n\cdot23\ldots(n-1)} r_{3n}\sigma_3\sigma_n\\
&\vdots\\
r_{1n}\sigma_1\sigma_n &= b_{12\cdot34\ldots n} r_{2n}\sigma_2\sigma_n + b_{13\cdot24\ldots n} r_{3n}\sigma_3\sigma_n + \ldots + b_{1n\cdot23\ldots(n-1)}\sigma_n^2
\end{aligned}\right]$$

...(10·36*b*)

Hence the eliminant of b's between (10·35) and (10·36*b*) is

$$\begin{vmatrix}
X_1 & X_2 & X_3 & \ldots & X_n\\
r_{12}\sigma_1\sigma_2 & \sigma_2^2 & r_{23}\sigma_2\sigma_3 & \ldots & r_{2n}\sigma_2\sigma_n\\
r_{13}\sigma_1\sigma_3 & r_{23}\sigma_2\sigma_3 & \sigma_3^2 & \ldots & r_{3n}\sigma_3\sigma_n\\
\vdots & \vdots & \vdots & & \vdots\\
r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & r_{3n}\sigma_3\sigma_n & \ldots & \sigma_n^2
\end{vmatrix} = 0 .$$

Dividing $C_1, C_2, \ldots, C_n$ by $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively and $R_2, R_3, \ldots R_n$ by $\sigma_2, \sigma_3, \ldots, \sigma_n$ respectively, we get

$$\begin{vmatrix}
\dfrac{X_1}{\sigma_1} & \dfrac{X_2}{\sigma_2} & \dfrac{X_3}{\sigma_3} & \ldots & \dfrac{X_n}{\sigma_n}\\
r_{12} & 1 & r_{32} & \ldots & r_{2n}\\
r_{13} & r_{23} & 1 & \ldots & r_{3n}\\
\vdots & \vdots & \vdots & & \vdots\\
r_{1n} & r_{2n} & r_{3n} & & 1
\end{vmatrix} = 0$$

...(10·37)

If we write

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 1 & \cdots & r_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{vmatrix} \quad \ldots(10\cdot38)$$

and ω_{ij} is the cofactor of the element in the ith row and jth column of ω, we get from (10·37)

$$\frac{X_1}{\sigma_1} \cdot \omega_{11} + \frac{X_2}{\sigma_2} \omega_{12} + \frac{X_3}{\sigma_3} \omega_{13} + \ldots + \frac{X_n}{\sigma_n} \omega_{1n} = 0 \quad \ldots(10\cdot39)$$

as the required equation of the plane of regression of X_1 on $X_2, X_3, \ldots, X_n$.

Equation (10·39) can be re-written as

$$X_1 = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} X_2 - \frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} X_3 - \ldots - \frac{\sigma_1}{\sigma_n} \cdot \frac{\omega_{1n}}{\omega_{11}} X_n \quad \ldots(10\cdot39a)$$

Comparing (10·39a) with (10·35), we get

$$\left.\begin{aligned} b_{12\cdot34\ldots n} &= -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \\ b_{13\cdot24\ldots n} &= -\frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} \\ &\vdots \\ b_{1n\cdot23\ldots(n-1)} &= -\frac{\sigma_1}{\sigma_n} \cdot \frac{\omega_{1n}}{\omega_{11}} \end{aligned}\right\} \quad \ldots(10\cdot40)$$

Remarks 1. From the symmetry of the result obtained in (10·40), the equation of the plane of regression of X_i, (say), on the remaining variables X_j $(j \neq i = 1, 2, \ldots, n)$, is given by

$$\frac{X_1}{\sigma_1} \omega_{i1} + \frac{X_2}{\sigma_2} \omega_{i2} + \ldots + \frac{X_i}{\sigma_i} \omega_{ii} + \ldots + \frac{X_n}{\sigma_n} \omega_{in} = 0 \; ; \; i = 1, 2, , n \quad \ldots(10\cdot41)$$

2. We have

$$b_{12\cdot34\ldots n} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}}$$

and $$b_{21\cdot34\ldots n} = -\frac{\sigma_2}{\sigma_1} \frac{\omega_{21}}{\omega_{22}}$$

Since each of σ_1, σ_2, ω_{11} and ω_{22} is non-negative and $\omega_{12} = \omega_{21}$, [*c.f.* Remarks 3 and 4 to §10·14, page 10·113], the sign of each regression coefficient $b_{12\cdot34\ldots n}$ and $b_{21\cdot34\ldots n}$ depends on ω_{12}.

10·13. Properties of residuals

Property 1. *The sum of the product of any residual of order zero with any other residual of higher order is zero, provided the subscript of the former occurs among the secondary subscripts of the latter.*

The normal equations for estimating b's in trivariate and n-variate distributions, as obtained in equations (10·30a) and (10·36a), are

$$\Sigma X_2 X_{1\cdot 23} = 0, \Sigma X_3 X_{1\cdot 23} = 0$$

and $$\Sigma X_i X_{1\cdot 23\ldots n} = 0 \; ; \; i = 2, 3, \ldots, n$$

respectively. Here X_i, $(i = 1, 2, 3, \ldots, n)$ can be regarded as a residual of order zero. Hence the result.

Property 2. *The sum of the product of any two residuals in which all the secondary subscripts of the first occur among the secondary subscripts of the second is unaltered if we omit any or all of the secondary subscripts of the first. Conversely, the product sum of any residual of order 'p' with a residual of order $p + q$, the 'p' subscripts being the same in each case is unaltered by adding to the secondary subscripts of the former any or all the 'q' additional subscripts of the latter.*

Let us consider

$$\begin{aligned}\Sigma X_{1\cdot 2} X_{1\cdot 23} &= \Sigma (X_1 - b_{12} X_2) X_{1\cdot 23} = \Sigma X_1 X_{1\cdot 23} - b_{12} \Sigma X_2 X_{1\cdot 23} \\ &= \Sigma X_1 X_{1\cdot 23} \qquad (c.f. \text{ Property 1})\end{aligned}$$

Also
$$\begin{aligned}\Sigma X_{1\cdot 23}^2 &= \Sigma X_{1\cdot 23} X_{1\cdot 23} = \Sigma (X_1 - b_{12\cdot 3} X_2 - b_{13\cdot 2} X_3) X_{1\cdot 23} \\ &= \Sigma X_1 X_{1\cdot 23} - b_{12\cdot 3} \Sigma X_2 X_{1\cdot 23} - b_{13\cdot 2} \Sigma X_3 X_{1\cdot 23} \\ &= \Sigma X_1 X_{1\cdot 23} \qquad (c.f. \text{ Property 1})\end{aligned}$$

$$\therefore \quad \Sigma X_{1\cdot 23}^2 = \Sigma X_{1\cdot 2} X_{1\cdot 23} = \Sigma X_1 X_{1\cdot 23}$$

Again $\Sigma X_{1\cdot 34\ldots n} X_{2\cdot 34\ldots n}$

$$= \Sigma[(X_1 - b_{13\cdot 4\ldots n} X_3 - b_{14\cdot 35\ldots n} X_4 - \ldots - b_{1n\cdot 34\ldots(n-1)} X_n) X_{2\cdot 34\ldots n}]$$

$$= \Sigma X_1 X_{2\cdot 34\ldots n} \qquad (c.f. \text{ Property 1})$$

Hence the property 2.

Property 3. *The sum of the product of two residuals is zero if all the subscripts (primary as well as secondary) of the one occur among the secondary subscripts of the other, e.g.,*

$$\Sigma X_{1\cdot 2} X_{3\cdot 12} = \Sigma (X_1 - b_{12} X_2) X_{3\cdot 12} = \Sigma X_1 X_{3\cdot 12} - b_{12} \Sigma X_2 X_{3\cdot 12} = 0 \qquad (c.f. \text{ Property 1})$$

$\Sigma X_{2\cdot 34\ldots n} X_{1\cdot 23\ldots n}$

$$= \Sigma[(X_2 - b_{23\cdot 4\ldots n} X_3 - b_{24\cdot 35\ldots n} X_4 - \ldots - b_{2n\cdot 34\ldots(n-1)} X_n) X_{1\cdot 23\ldots n}]$$

$$\begin{aligned}&= \Sigma X_2 X_{1\cdot 23\ldots n} - b_{23\cdot 4\ldots n} \Sigma X_3 X_{1\cdot 23\ldots n} - b_{24\cdot 35\ldots n} \Sigma X_4 X_{1\cdot 23\ldots n} \\ &\qquad \ldots - b_{2n\cdot 34\ldots(n-1)} \Sigma X_n X_{1\cdot 23\ldots n}\end{aligned}$$

$$= 0 \qquad (c.f. \text{ Property 1})$$

Hence the property 3.

10·13·1. Variance of the Residual. Let us consider the plane of regression of X_1 on $X_2, X_3, \ldots, X_n$ *viz.*,

$$X_1 = b_{12\cdot34\ldots n} X_2 + b_{13\cdot24\ldots n} X_3 + \ldots + b_{1n\cdot23\ldots(n-1)} X_n$$

Since all the X_i's are measured from their respective means, we have

$$E(X_i) = 0;\ i = 1, 2, \ldots, n \Rightarrow E(X_{1\cdot23\ldots n}) = 0$$

Hence the variance of the residual is given by

$$\sigma^2_{1\cdot23\ldots n} = \frac{1}{N}\Sigma[X_{1\cdot23\ldots n} - E(X_{1\cdot23\ldots n})]^2 = \frac{1}{N}\Sigma X^2_{1\cdot23\ldots n}$$

$$= \frac{1}{N}\Sigma X_{1\cdot23\ldots n} X_{1\cdot23\ldots n} = \frac{1}{N}\Sigma X_1 X_{1\cdot23\ldots n},$$

(*c.f.* Property 2 § 10·13)

$$= \frac{1}{N}\Sigma X_1 (X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n)$$

$$= \sigma_1^2 - b_{12\cdot34\ldots n} r_{12}\sigma_1\sigma_2 - b_{13\cdot24\ldots n} r_{13}\sigma_1\sigma_3 - \ldots - b_{1n\cdot23\ldots(n-1)} r_{1n}\sigma_1\sigma_n$$

$$\Rightarrow \quad \sigma_1^2 - \sigma^2_{1\cdot23\ldots n} = b_{12\cdot34\ldots n} r_{12}\sigma_1\sigma_2 - b_{13\cdot24\ldots n} r_{13}\sigma_1\sigma_3 - \ldots - b_{1n\cdot23\ldots(n-1)} r_{1n}\sigma_1\sigma_n \quad \ldots(10\cdot42)$$

Eliminating the b's in equations (10·42) and (10·36b), we get

$$\begin{vmatrix} \sigma_1^2 - \sigma^2_{1\cdot23\ldots n} & r_{12}\sigma_1\sigma_2 & \ldots & r_{1n}\sigma_1\sigma_n \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & \ldots & r_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & & \vdots \\ r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & \ldots & \sigma_n^2 \end{vmatrix} = 0$$

Dividing $R_1, R_2, \ldots, R_n$, by $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively and also $C_1, C_2, \ldots, C_n$ by $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively, we get

$$\begin{vmatrix} 1 - \dfrac{\sigma^2_{1\cdot23\ldots n}}{\sigma_1^2} & r_{12} & \ldots & r_{1n} \\ r_{12} & 1 & \ldots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{1n} & r_{2n} & \ldots & 1 \end{vmatrix} = 0$$

$$\Rightarrow \quad \begin{vmatrix} 1 - \dfrac{\sigma^2_{1\cdot23\ldots n}}{\sigma_1^2} & r_{12} & \ldots & r_{1n} \\ r_{12} + 0 & 1 & \ldots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{1n} + 0 & r_{2n} & \ldots & 1 \end{vmatrix} = 0$$

$$\begin{vmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{12} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{1n} & r_{2n} & \cdots & 1 \end{vmatrix} - \begin{vmatrix} \dfrac{\sigma^2_{1\cdot 23\ldots n}}{\sigma_1^2} & r_{12} & \cdots & r_{1n} \\ 0 & 1 & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & r_{2n} & \cdots & 1 \end{vmatrix} = 0$$

$$\Rightarrow \qquad \omega - \frac{\sigma^2_{1\cdot 23\ldots n}}{\sigma_1^2}\,\omega_{11} = 0$$

$$\therefore \qquad \sigma^2_{1\cdot 23\ldots n} = \sigma_1^2\,\frac{\omega}{\omega_{11}} \qquad \ldots(10\cdot 43)$$

Remark. In a tri-variate distribution,

$$\sigma_{1\cdot 23}^2 = \sigma_1^2\,\frac{\omega}{\omega_{11}} \qquad \ldots(10\cdot 43a)$$

where ω and ω_{11} are defined in (10·32).

10·14. Coefficient of Multiple Correlation. In a tri-variate distribution in which each of the variables X_1, X_2, and X_3 has N observations, the multiple correlation coefficient of X_1 on X_2 and X_3, usually denoted by $R_{1\cdot 23}$, is the simple correlation coefficient between X_1 and the joint effect of X_2 and X_3 on X_1. In other words $R_{1\cdot 23}$ is the correlation coefficient between X_1 and its estimated value as given by the plane of regression of X_1 on X_2 and X_3 *viz.*,

$$e_{1\cdot 23} = b_{12\cdot 3}X_2 + b_{13\cdot 2}X_3$$

We have $\quad X_{1\cdot 23} = X_1 - b_{12\cdot 3}X_2 - b_{13\cdot 2}X_3 = X_1 - e_{1\cdot 23}$

$$\Rightarrow \qquad e_{1\cdot 23} = X_1 - X_{1\cdot 23}$$

Since X_i's are measured from their respective means, we have

$$E(X_{1\cdot 23}) = 0 \text{ and } E(e_{1\cdot 23}) = 0 \qquad (\because E(X_i) = 0;\ i = 1, 2, 3)$$

By def.,

$$R_{1\cdot 23} = \frac{\text{Cov}(X_1, e_{1\cdot 23})}{\sqrt{V(X_1)\,V(e_{1\cdot 23})}}. \qquad \ldots(10\cdot 44)$$

$$\begin{aligned}
\text{Cov}(X_1, e_{1\cdot 23}) &= E[\{X_1 - E(X_1)\}\{e_{1\cdot 23} - E(e_{1\cdot 23})\}] = E(X_1\, e_{1\cdot 23}) \\
&= \frac{1}{N}\sum X_1\, e_{1\cdot 23} = \frac{1}{N}\sum X_1\,(X_1 - X_{1\cdot 23}) \\
&= \frac{1}{N}\sum X_1^2 - \frac{1}{N}\sum X_1X_{1\cdot 23} = \frac{1}{N}\sum X_1^2 - \frac{1}{N}\sum X^2_{1\cdot 23} \\
&= \sigma_1^2 - \sigma_{1\cdot 23}^2 \qquad (c.f.\ \text{Property 2, § 10·13})
\end{aligned}$$

$$\begin{aligned}
\text{Also} \quad V(e_{1\cdot 23}) &= E(e_{1\cdot 23}^2) = \frac{1}{N}\sum e_{1\cdot 23}^2 = \frac{1}{N}\sum (X_1 - X_{1\cdot 23})^2 \\
&= \frac{1}{N}\sum (X_1^2 + X_{1\cdot 23}^2 - 2\,X_1X_{1\cdot 23}) \\
&= \frac{1}{N}\sum X_1^2 + \frac{1}{N}\sum X_{1\cdot 23}^2 - \frac{2}{N}\sum X_1X_{1\cdot 23}
\end{aligned}$$

$$= \frac{1}{N} \Sigma X_1^2 + \frac{1}{N} \Sigma X_{1\cdot 23}^2 - \frac{2}{N} \Sigma X_{1\cdot 23}^2$$

$$= \sigma_1^2 - \sigma_{1\cdot 23}^2 \qquad (c.f. \text{ Property 2, § 10·13})$$

$$\therefore \qquad R_{1\cdot 23} = \frac{\sigma_1^2 - \sigma_{1\cdot 23}^2}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma_{1\cdot 23}^2)}}$$

$$\Rightarrow \qquad R^2_{1\cdot 23} = \frac{\sigma_1^2 - \sigma_{1\cdot 23}^2}{\sigma_1^2} = 1 - \frac{\sigma_{1\cdot 23}^2}{\sigma_1^2}$$

$$\Rightarrow \qquad 1 - R^2_{1\cdot 23} = \frac{\sigma_{1\cdot 23}^2}{\sigma_1^2}$$

Using (10·43a), we get

$$1 - R^2_{1\cdot 23} = \frac{\omega}{\omega_{11}} \qquad \ldots(10\cdot 45)$$

where

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} \text{ (On simplification)}.$$

and

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2$$

Hence from (10·45), we get

$$R^2_{1\cdot 23} = 1 - \frac{\omega}{\omega_{11}} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2} \qquad \ldots(10\cdot 45a)$$

This formula expresses the multiple correlation coefficient in terms of the total correlation coefficients between the pairs of variables.

Generalisation. In case of n-variate distribution, the multiple correlation coefficient of X_1 on $X_2, X_3, \ldots, X_n$, usually denoted by $R_{1\cdot 23\ldots n}$, is the correlation coefficient between X_1 and

$$e_{1\cdot 23\ldots n} = X_1 - X_{1\cdot 23\ldots n}$$

$$\therefore \qquad R_{1\cdot 23\ldots n} = \frac{\text{Cov}(X_1, e_{1\cdot 23\ldots n})}{\sqrt{V(X_1)\, V(e_{1\cdot 23\ldots n})}}$$

$$\text{Cov}(X_1, e_{1\cdot 23\ldots n}) = \frac{1}{N} \Sigma X_1\, e_{1\cdot 23\ldots n} = \frac{1}{N} \Sigma X_1(X_1 - X_{1\cdot 23\ldots n})$$

$$= \frac{1}{N} \Sigma X_1^2 - \frac{1}{N} \Sigma X_1 X_{1\cdot 23\ldots n}$$

$$= \frac{1}{N} \Sigma X_1^2 - \frac{1}{N} \Sigma X^2_{1\cdot 22\ldots n} = \sigma_1^2 - \sigma^2_{1\cdot 23\ldots n} \qquad \ldots(*)$$

$$V(e_{1\cdot 23\ldots n}) = \frac{1}{N} \Sigma e^2_{1\cdot 23\ldots n} = \frac{1}{N} \Sigma (X_1 - X_{1\cdot 23\ldots n})^2$$

$$= \frac{1}{N} \Sigma \left(X_1^2 + X^2_{1\cdot 23\ldots n} - 2X_1 X_{1\cdot 23\ldots n}\right)$$

$$= \frac{1}{N} \Sigma X_1^2 + \frac{1}{N} \Sigma X^2_{1\cdot 23\ldots n} - 2 \frac{1}{N} \Sigma X_1 X_{1\cdot 23\ldots n}$$

$$= \frac{1}{N} \Sigma X_1^2 + \frac{1}{N} \Sigma X^2_{1\cdot 23\ldots n} - \frac{2}{N} \Sigma X^2_{1\cdot 23\ldots n}$$

$$= \sigma_1^2 - \sigma^2_{1\cdot 23\ldots n}$$

$$\therefore \quad R_{1\cdot 23\ldots n} = \frac{\sigma_1^2 - \sigma^2_{1\cdot 23\ldots n}}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma^2_{1\cdot 23\ldots n})}} = \left(\frac{\sigma_1^2 - \sigma^2_{1\cdot 23\ldots n}}{\sigma_1^2}\right)^{1/2}$$

$$R^2_{1\cdot 23\ldots n} = 1 - \frac{\sigma^2_{1\cdot 23\ldots n}}{\sigma_1^2} = 1 - \frac{\omega}{\omega_{11}} \qquad \ldots(10\cdot 45c)$$

where ω and ω_{11} are defined in (10·38).

Remarks 1. It may be pointed out here that multiple correlation coefficient can never be negative, because from (*) and (**), we get

$$\text{Cov}\,(X_1, e_{1\cdot 23\ldots n}) = \sigma_1^2 - \sigma^2_{1\cdot 23\ldots n} = \text{Var}\,(e_{1\cdot 23\ldots n}) \geq 0$$

Since the sign of $R_{1\cdot 23\ldots n}$ depends upon the covariance term $\text{Cov}\,(X_1, e_{1\cdot 23\ldots n})$, we conclude that $R_{1\cdot 23\ldots n} \geq 0$.

2. Since $R^2_{1\cdot 23\ldots n} \geq 0$, we have :

$$1 - \frac{\omega}{\omega_{11}} \geq 0 \quad \Rightarrow \quad \omega \leq \omega_{11} \qquad \ldots(10\cdot 45d)$$

3. Also, $R^2_{1\cdot 23\ldots n} \leq 1 \quad \Rightarrow \quad 1 - \dfrac{\omega}{\omega_{11}} \leq 1$

$$\Rightarrow \quad 0 \leq \frac{\omega}{\omega_{11}} \quad \Rightarrow \quad \frac{\omega}{\omega_{11}} \geq 0 \quad \Rightarrow \quad \omega \geq 0 \qquad \ldots(10\cdot 45e)$$

From the above results, we get

$$\omega_{11} \geq \omega \geq 0 \qquad \ldots(10\cdot 45f)$$

In general, we have

$$\omega_{ii} \geq 0 \;;\; i = 1, 2, \ldots, n$$

4. Since ω is symmetric in r_{ij}'s, we have

$$\omega_{ij} = \omega_{ji} \;;\; i \neq j = 1, 2, \ldots, n \qquad \ldots(10\cdot 45g)$$

10·14·1. Properties of Multiple Correlation Coefficient

1. Multiple correlation co-efficient measures the closeness of the association between the observed values and the expected values of a variable obtained from the multiple linear regression of that variable on other variables.

2. Multiple correlation coefficient between observed values and expected values, when the expected values are calculated from a linear relation of the variables determined by the method of least squares, is always greater than that where expected values are calculated from any other linear combination of the variables.

3. Since $R_{1.23}$ is the simple correlation between X_1 and $e_{1.23}$, it must lie between -1 and $+1$. But as seen in Remark 1 above, $R_{1.23}$ is a non-negative quantity and we conclude that $0 \le R_{1.23} \le 1$.

4. If $R_{1.23} = 1$, then association is perfect and all the regression residuals are zero, and as such $\sigma^2_{1.23} = 0$. In ths case, since $X_1 = e_{1.23}$, the predicted value of X_1, the multiple linear regression equation of X_1 on X_2 and X_3 may be said to be a perfect prediction formula.

5. If $R_{1.23} = 0$, then all total and partial correlations involving X_1 are zero. [See Example 10·37). So X_1 is completely uncorrelated with all the other variables in this case and the multiple regression equation fails to throw any light on the value of X_1 when X_2 and X_3 are known.

6. $R_{1.23}$ is not less than any total correlation coefficient, *i.e.*,

$$R_{1.23} \ge r_{12}, r_{13}, r_{23}$$

10·15. Coefficient of Partial Correlation. Sometimes the correlation between two variables X_1 and X_2 may be partly due to the correlation of a third variable, X_3 with both X_1 and X_2. In such a situation, one may want to know what the correlation between X_1 and X_2 would be if the effect of X_3 on each of X_1 and X_2 were eliminated. This correlation is called the *partial correlation* and the correlation coefficient between X_1 and X_2 after the linear effect of X_3 on each of them has been eliminated is called the *partial correlation coefficient.*

The residual $X_{1.3} = X_1 - b_{13}X_3$, may be regarded as that part of the variable X_1 which remains after the linear effect of X_3 has been eliminated. Similarly, the residual $X_{2.3}$ may be interpreted as the part of the variable X_2 obtained after eliminating the linear effect of X_3. Thus the partial correlation coefficient between X_1 and X_2, usually denoted by $r_{12.3}$, is given by

$$r_{12.3} = \frac{\text{Cov}(X_{1.3}, X_{2.3})}{\sqrt{\text{Var}(X_{1.3})\,\text{Var}(X_{2.3})}} \qquad \ldots(10·46)$$

We have

$$\begin{aligned}
\text{Cov}(X_{1.3}, X_{2.3}) &= \frac{1}{N}\Sigma X_{1.3}X_{2.3} = \frac{1}{N}\Sigma X_1X_{2.3}\\
&= \frac{1}{N}\Sigma X_1(X_2 - b_{23}X_3) = \frac{1}{N}\Sigma X_1X_2 - b_{23}\frac{1}{N}\Sigma X_1X_3\\
&= r_{12}\sigma_1\sigma_2 - r_{23}\frac{\sigma_2}{\sigma_3}\cdot(r_{13}\sigma_1\sigma_3).\\
&= \sigma_1\sigma_2(r_{12} - r_{13}r_{23})
\end{aligned}$$

Also

$$\begin{aligned}
V(X_{1.3}) &= \frac{1}{N}\Sigma X_{1.3}^2 = \frac{1}{N}\Sigma X_{1.3}X_{1.3}\\
&= \frac{1}{N}\Sigma X_1X_{1.3} = \frac{1}{N}\Sigma X_1(X_1 - b_{13}X_3)
\end{aligned}$$

$$= \frac{1}{N} \Sigma X_1^2 - b_{13} \cdot \frac{1}{N} \Sigma X_1 X_3$$

$$= \sigma_1^2 - r_{13} \frac{\sigma_1}{\sigma_3} r_{13} \sigma_1 \sigma_3$$

$$= \sigma_1^2 (1 - r_{13}^2)$$

Similarly, we shall get

$$V(X_{2\cdot3}) = \sigma_2^2 (1 - r_{23}^2)$$

Hence

$$r_{12\cdot3} = \frac{\sigma_1 \sigma_2 (r_{12} - r_{13} r_{23})}{\sqrt{\sigma_1^2 (1 - r_{13}^2)\, \sigma_2^2 (1 - r_{23}^2)}} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad \text{...(10·46a)}$$

Aliter. We have

$$0 = \Sigma X_{2\cdot3} X_{1\cdot23}$$
$$= \Sigma X_{2\cdot3} (X_1 - b_{12\cdot3} X_2 - b_{13\cdot2} X_3)$$
$$= \Sigma X_1 X_{2\cdot3} - b_{12\cdot3} \Sigma X_{2\cdot3} X_2 - b_{13\cdot2} \Sigma X_{2\cdot3} X_3$$
$$= \Sigma X_{1\cdot3} X_{2\cdot3} - b_{12\cdot3} \Sigma X_{2\cdot3} X_{2\cdot3}$$

$$\therefore \qquad b_{12\cdot3} = \frac{\Sigma X_{1\cdot3} X_{2\cdot3}}{\Sigma X_{2\cdot3}^2}$$

From this it follows that $b_{12\cdot3}$ is coefficient of regression of $X_{1\cdot3}$ on $X_{2\cdot3}$.

Similarly, $b_{21\cdot3}$ is the coefficient of regression of $X_{2\cdot3}$ on $X_{1\cdot3}$.

Since correlation coefficient is the geometric mean between regression coefficients, we have

$$r^2_{12\cdot3} = b_{12\cdot3} \times b_{21\cdot3}$$

But by def.,

$$b_{12\cdot3} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \quad \text{and} \quad b_{21\cdot3} = -\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}}$$

$$\therefore \qquad r^2_{12\cdot3} = \left(-\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}}\right)\left(-\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}}\right) = \frac{\omega_{12}^2}{\omega_{11}\,\omega_{22}}$$

$$(\because \omega_{12} = \omega_{21})$$

$$\Rightarrow \qquad r_{12\cdot3} = -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}},$$

the negative sign being taken since the sign of regression coefficients is the same as that of $(-\omega_{12})$.

Substituting the values of ω_{12}, ω_{11} and ω_{22} from (10·32), we get

$$r_{12\cdot3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Remarks 1. The expressions for $r_{13\cdot2}$ and $r_{23\cdot1}$ can be similarly obtained, to give

$$r_{13\cdot 2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - r_{12}^2)\,(1 - r_{32}^2)}} \quad \text{and} \quad r_{23\cdot 1} = \frac{r_{23} - r_{21} r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}}$$

2. If $r_{12\cdot 3} = 0$, we have then $r_{12} = r_{13}\, r_{23}$, it means that r_{12} will not be zero if X_3 is correlated with both X_1 and X_2. Thus, although X_1 and X_2 may be uncorrelated when effect of X_3 is eliminated, yet X_1 and X_2 may appear to be correlated because they carry the effect of X_3 on them.

3. Partial correlation coefficient helps in deciding whether to include or not an additional independent variable in regression analysis.

4. We know that $\sigma_1^2(1 - r_{12}^2)$ and $\sigma_1^2(1 - r_{13}^2)$ are the residual variances if X_1 is estimated from X_2 and X_3 individually, while $\sigma_1^2\,(1 - R_{1\cdot 23}^2)$ is the residual variance if X_1 is estimated from X_2 and X_3 taken together. So from the above remark and $R_{1\cdot 23}^2 \geq r_{12}^2$ and r_{13}^2, it follows that inclusion of an additional variable can only reduce the residual variance. Now inclusion of X_3 when X_2 has already been taken for predicting X_1, is worthwhile only when the resultant reduction in the residual variance is substantial. This will be the case when $r_{13\cdot 2}$ is sufficiently large. Thus in this respect partial correlation coefficient has its significance in regression analysis.

10·15·1. Generalisation. In the case of n variables $X_1, X_2, \ldots, X_n$ the partial correlation coefficient $r_{12\cdot 34\ldots n}$ between X_1 and X_2 (after the linear effect of $X_3, X_4, \ldots, X_n$ on them has been eliminated), is given by

$$r^2_{12\cdot 34\ldots n} = b_{12\cdot 34\ldots n} \times b_{21\cdot 34\ldots n}$$

But, we have

$$\left.\begin{aligned} b_{12\cdot 34\ldots n} &= -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}} \\ \text{and} \qquad b_{21\cdot 34\ldots n} &= -\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}} \end{aligned}\right\} \text{[c.f. Equation (10·40)].}$$

$$\therefore \qquad r^2_{12\cdot 34\ldots n} = \left(-\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}}\right)\left(-\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}}\right) = \frac{\omega_{12}^2}{\omega_{11}\omega_{22}}$$

$$\Rightarrow \qquad r_{12\cdot 34\ldots n} = -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}} \qquad (10\text{·}46b)$$

negative sign being taken since the sign of the regression coefficient is same as that of $(-\omega_{12})$.

10·16. Multiple Correlation in Terms of Total and Partial Correlations.

$$1 - R_{1\cdot 23}^2 = (1 - r_{12}^2)(1 - r_{13\cdot 2}^2) \qquad \ldots(10\text{·}46c)$$

Proof. We have

$$\begin{aligned} 1 - R_{1\cdot 23}^2 &= 1 - \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2} \\ &= \frac{1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2} \end{aligned}$$

Also

$$1 - r_{13\cdot2}^2 = 1 - \frac{(r_{13} - r_{12}\, r_{23})^2}{(1 - r_{12}^2)(1 - r_{23}^2)} = \frac{1 - r_{12}^2 - r_{23}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23}}{(1 - r_{12}^2)(1 - r_{23}^2)}$$

Hence the result.

Theorem. *Any standard deviation of order 'p' may be expressed in terms of a standard deviation of order (p – 1) and a partial correlation coefficient of order (p – 1).*

Proof. Let us consider the sum :

$$\begin{aligned}
\Sigma X^2_{1\cdot23\ldots n} &= \Sigma X_{1\cdot23\ldots n} X_{1\cdot23\ldots n} \\
&= \Sigma [X_{1\cdot23\ldots(n-1)} X_{1\cdot23\ldots n}] \\
& \qquad\qquad\qquad (c.f.\ \text{Property 2, § 10·13}) \\
&= \Sigma [X_{1\cdot23\ldots(n-1)} (X_1 - b_{12\cdot34\ldots n} X_2 - \ldots - b_{1(n-1).23\ldots n} X_{n-1} \\
& \qquad\qquad - b_{1n\cdot23\ldots(n-1)} X_n)] \\
&= \Sigma X_{1\cdot23\ldots(n-1)} X_1 - b_{1n.23\ldots(n-1)} \Sigma X_{1\cdot23\ldots(n-1)} X_n \\
& \qquad\qquad\qquad (c.f.\ \text{Property 2 § 10·13}) \\
&= \Sigma X^2_{1\cdot23\ldots(n-1)} - b_{1n\,23\ldots(n-1)} \Sigma X_{1\cdot23\ldots(n-1)} X_{n\cdot23\ldots(n-1)}
\end{aligned}$$

Dividing both sides by N (total number of observations), we get $\sigma^2_{1\cdot23\ldots n} = \sigma^2_{1\,23\ldots(n-1)} - b_{1n\cdot23\ldots(n-1)}\ \text{Cov}\ (X_{1\,23\ldots(n-1)}, X_{n\cdot23\ldots(n-1)})$

The regression coefficient of $X_{n\,23\ldots(n-1)}$ on $X_{1\cdot23\ldots(n-1)}$ is given by

$$b_{n1\cdot23\ldots(n-1)} = \frac{\text{Cov}\ (X_{1\cdot23\ldots(n-1)}, X_{n\cdot23\ldots(n-1)})}{\sigma^2_{1\cdot23\ldots(n-1)}}$$

$$\begin{aligned}
\therefore \qquad \sigma^2_{1\cdot23\ldots n} &= \sigma^2_{1\cdot23\ldots(n-1)} [1 - b_{1n\cdot23\ldots(n-1)} \cdot b_{n1\cdot23\ldots(n-1)}] \\
&= \sigma^2_{1\cdot23\ldots(n-1)} [1 - r^2_{n1\,23\ldots(n-1)}], \qquad \ldots(10\cdot47)
\end{aligned}$$

a formula which expresses the standard deviation of order $(n - 1)$ in terms of standard deviation of order $(n - 2)$ and partial correlation coefficient of order $(n - 2)$. If we take $p = (n - 1)$, the theorem is established.

Cor. 1. From (10·47), we have

$$\sigma^2_{1\,23\ldots(n-1)} = \sigma^2_{1\cdot23\ldots(n-2)} (1 - r^2_{1(n-1)\cdot23\ldots(n-2)}) \qquad \ldots(10\cdot47a)$$

and so on. Thus the repeated application of (10·47) gives

$$\sigma^2_{1\cdot23\ldots n} = \sigma_1^2 (1 - r_{12}^2)(1 - r_{13\cdot2}^2)(1 - r_{14\cdot32}^2)\ldots(1 - r^2_{1n\cdot23\ldots(n-1)}) \qquad \ldots(10\cdot47b)$$

Since partial correlation coefficients cannot exceed unity numerically, we get from (10·47), (10·47a), and so on,

$$\left.\begin{array}{ccc} \sigma^2_{1\cdot 23\ldots n} & \le & \sigma^2_{1\cdot 23\ldots(n-1)} \\ \sigma^2_{1\cdot 23\ldots(n-1)} & \le & \sigma^2_{1\cdot 23\ldots(n-2)} \\ | & & | \\ \sigma_{1\cdot 23}^2 & \le & \sigma_{1\cdot 2}^2 \\ \sigma_{1\cdot 2}^2 & \le & \sigma_1^2 \end{array}\right\}$$

$\Rightarrow$ $$\sigma_1 \ge \sigma_{1\cdot 2} \ge \sigma_{1\cdot 23} \ge \ldots \ge \sigma_{1\cdot 23\ldots n} \qquad \ldots(10\cdot 47c)$$

Cor. 2. Also, we have

$$\sigma^2_{1\cdot 23\ldots n} = \sigma_1^2(1 - R^2_{1\cdot 23\ldots n})$$

On using (10·47*b*), we get

$$1 - R^2_{1\cdot 23\ldots n} = (1 - r_{12}^2)(1 - r_{13\cdot 2}^2)\ldots(1 - r^2_{1n\cdot 3\ldots(n-1)}) \qquad \ldots(10\cdot 47d)$$

This is the generalisation of the result obtained in (10·46*c*).

Since $\quad | r_{ij.(s)} | \le 1;\ s = 0, 1, 2, \ldots, (n-1),$

where $r_{ij.(s)}$ is a partial correlation coefficient of order s. we get from (10·47*d*)

$$1 - R^2_{1\cdot 23\ldots n} \le 1 - r_{12}^2$$
$$1 - R^2_{1\cdot 23\ldots n} \le 1 - r^2_{13\cdot 2},$$

and so on.

i.e., $$R^2_{1\cdot 23\ldots n} \ge r_{12}^2, r^2_{13\cdot 2}, \ldots, r^2_{1n\cdot 23\ldots(n-1)} \qquad \ldots(10\cdot 47e)$$

Since $R_{1\cdot 23\ldots n}$ is symmetric in its secondary subscripts, we have

$$\left.\begin{array}{l} R^2_{1\cdot 23\ldots n} \ge r_{1i}^2\ ,\ (i = 2, 3, \ldots, n) \\ R^2_{1\cdot 23\ldots n} \ge r_{1i.j}\ (i \ne j = 2, 3, \ldots, n) \end{array}\right\} \qquad \ldots(10\cdot 47f)$$

and so on

10·17. Expression for Regression Coefficients in Terms of Regression Coefficients of Lower Order. Consider

$$\begin{aligned} \Sigma X_{1\cdot 34\ldots n} X_{2\cdot 34\ldots n} &= \Sigma X_{1\cdot 34\ldots(n-1)} X_{2\cdot 34\ldots n} \\ &= \Sigma X_{1\cdot 34\ldots(n-1)}(X_2 - b_{23\cdot 4\ldots n}X_3 - \ldots - b_{2n\cdot 34\ldots(n-1)}X_n) \\ &= \Sigma X_{1\cdot 34\ldots(n-1)} X_2 - b_{2n\cdot 34\ldots(n-1)} \Sigma X_{1\cdot 34\ldots(n-1)} X_n \\ &= \Sigma X_{1\cdot 34\ldots(n-1)} X_{2\cdot 34\ldots(n-1)} \\ &\qquad - b_{2n\cdot 34\ldots(n-1)} \Sigma X_{1\cdot 34\ldots(n-1)} X_{n\cdot 34\ldots(n-1)} \end{aligned}$$

Dividing both sides by N, the total number of observations, we get

$$\begin{aligned} \text{Cov}(X_{1\cdot 34\ldots n}, X_{2\cdot 34\ldots n}) &= \text{Cov}(X_{1\cdot 34\ldots(n-1)}, X_{2\cdot 34\ldots(n-1)}) \\ &\qquad - b_{2n.34\ldots(n-1)} \text{Cov}(X_{1\cdot 34\ldots(n-1)}, X_{n\cdot 34\ldots(n-1)}) \end{aligned}$$

$$\begin{aligned} b_{12\cdot 34\ldots n}\, \sigma^2_{2\cdot 34\ldots n} &= b_{12\cdot 34\ldots(n-1)}\, \sigma^2_{2\cdot 34\ldots(n-1)} \\ &\qquad - b_{2n.34\ldots(n-1)}\, b_{1n\cdot 34\ldots(n-1)}\, \sigma^2_{n\cdot 34\ldots(n-1)} \end{aligned}$$

On using (10·47), we get

$$b_{12\cdot34\ldots n}\,\sigma^2_{2\cdot34\ldots(n-1)}\,\{1 - r^2_{2n\cdot34\ldots(n-1)}\}$$
$$= \sigma^2_{2\cdot34\ldots(n-1)}\left[b_{12\cdot34\ldots(n-1)} - b_{2n\cdot34\ldots(n-1)}\,b_{1n\cdot34\ldots(n-1)} \times \frac{\sigma^2_{n\cdot34\ldots(n-1)}}{\sigma^2_{2\cdot34\ldots(n-1)}}\right] \quad \ldots(*)$$

In the case of two variables, we have

$$b_{ij}\,\sigma_j^2 = b_{ji}\,\sigma_i^2 = \text{Cov}\,(X_i, X_j) \quad \Rightarrow \quad b_{ij} = \frac{\sigma_i^2}{\sigma_j^2}\,b_{ji}$$

$$\therefore \qquad b_{2n\cdot34\ldots(n-1)}\;\frac{\sigma^2_{n\cdot34\ldots(n-1)}}{\sigma^2_{2\cdot34\ldots(n-1)}} = b_{n2\cdot34\ldots(n-1)}$$

Hence from (*), we get

$$b_{12\cdot34\ldots n}\,\sigma^2_{2\cdot34\ldots(n-1)}\,\{1 - r^2_{2n\cdot34\ldots(n-1)}\}$$
$$= \sigma^2_{2\cdot34\ldots(n-1)}\,[b_{12\cdot34\ldots(n-1)} - b_{1n\cdot34\ldots(n-1)}\,b_{n2\cdot34\ldots(n-1)}]$$

$$\therefore \quad b_{12\cdot34\ldots n} = \left[\frac{b_{12\cdot34\ldots(n-1)} - b_{1n\cdot34\ldots(n-1)}\,b_{n2\cdot34\ldots(n-1)}}{1 - r^2_{2n\cdot34\ldots(n-1)}}\right] \qquad \ldots(10\cdot48)$$

$$\Rightarrow \quad b_{12\cdot34\ldots n} = \left[\frac{b_{12\cdot34\ldots(n-1)} - b_{1n\cdot34\ldots(n-1)}\cdot b_{n2\cdot34\ldots(n-1)}}{1 - b_{2n\cdot34\ldots(n-1)}\cdot b_{n2\cdot34\ldots(n-1)}}\right] \qquad \ldots(10\cdot48a)$$

10·18. Expression for Partial Correlation Coefficient in Terms of Correlation Coefficients of Lower Order. By definition, we have

$$b_{ij.lm\ldots t} = r_{ij.lm\ldots t} \times \frac{\sigma_{i.lm\ldots t}}{\sigma_{j.lm\ldots t}} \qquad \ldots(*)$$

$$\therefore \qquad b_{1n\cdot34\ldots(n-1)}\cdot b_{n2\cdot34\ldots(n-1)}$$
$$= r_{1n\cdot34\ldots(n-1)}\,\frac{\sigma_{1\cdot34\ldots(n-1)}}{\sigma_{n\cdot34\ldots(n-1)}} \times r_{n2\cdot34\ldots(n-1)}\,\frac{\sigma_{n\cdot34\ldots(n-1)}}{\sigma_{2\cdot34\ldots(n-1)}}$$
$$= r_{1n\cdot34\ldots(n-1)}\cdot r_{n2\cdot34\ldots(n-1)}\cdot\frac{\sigma_{1.34\ldots(n-1)}}{\sigma_{2\cdot34\ldots(n-1)}} \qquad \ldots(**)$$

Hence from (10·48), on using (*) and (**), we get

$$r_{12\cdot34\ldots n} \times \frac{\sigma_{1\cdot34\ldots n}}{\sigma_{2\cdot34\ldots n}}$$
$$= \left[\frac{\{r_{12\cdot34\ldots(n-1)} - r_{1n.34\ldots(n-1)}\,r_{n2\cdot34\ldots(n-1)}\}}{1 - r^2_{2n\cdot34\ldots(n-1)}}\;\frac{\sigma_{1\cdot34\ldots(n-1)}}{\sigma_{2\cdot34\ldots(n-1)}}\right] \quad \ldots(***)$$

Also on using (10·47), we get

$$\frac{\sigma_{1\cdot34\ldots n}}{\sigma_{2\cdot34\ldots n}} = \frac{\sigma_{1\cdot34\ldots(n-1)}}{\sigma_{2\cdot34\ldots(n-1)}} \times \left[\frac{1 - r^2_{1n\cdot34\ldots(n-1)}}{1 - r^2_{2n\cdot34\ldots(n-1)}}\right]^{1/2}$$

Hence from (***), we get

$$r_{12\cdot34\ldots n}\left[\frac{1-r^2_{1n\cdot34\ldots(n-1)}}{1-r^2_{2n\cdot34\ldots(n-1)}}\right]^{\frac{1}{2}}$$

$$=\left[\frac{r_{12\cdot34\ldots(n-1)}-r_{1n\cdot34\ldots(n-1)}\,r_{n2\cdot34\ldots(n-1)}}{1-r^2_{2n\cdot34\ldots(n-1)}}\right]$$

$$\Rightarrow \qquad r_{12\cdot34\ldots n}=\frac{r_{12\cdot34\ldots(n-1)}-r_{1n\cdot34\ldots(n-1)}\,r_{n2\cdot34\ldots(n-1)}}{(1-r^2_{1n\cdot34\ldots(n-1)})^{1/2}\,(1-r^2_{n2\cdot34\ldots(n-1)})^{1/2}} \qquad \ldots(10\cdot49)$$

which is an expression for the correlation coefficient of order $p=(n-2)$ in terms of the correlation coefficient of order $(p-1)=(n-3)$.

Example 10·33. *From the data relating to the yield of dry bark* (X_1), *height* (X_2) *and girth* X_3 *for 18 cinchona plants the following correlation coefficients were obtained :*

$$r_{12}=0{\cdot}77,\; r_{13}=0{\cdot}72 \text{ and } r_{23}=0{\cdot}52$$

Find the the partial correlation coefficient $r_{12\cdot3}$ *and multiple correlation coefficient* $R_{1\cdot23}$.

Solution.

$$r_{12\cdot3}=\frac{r_{12}-r_{13}\,r_{23}}{\sqrt{(1-r_{13}^2)\,(1-r_{23}^2)}}=\frac{0{\cdot}77-0{\cdot}72\times0{\cdot}52}{\sqrt{[1-(0{\cdot}72)^2][1-(0{\cdot}52)^2]}}=0{\cdot}62$$

$$R_{1\cdot23}^2=\frac{r_{12}^2+r_{13}^2-2r_{12}\,r_{13}\,r_{23}}{1-r_{23}^2}$$

$$=\frac{(0{\cdot}77)^2+(0{\cdot}72)^2-2\times0{\cdot}77\times0{\cdot}72\times0{\cdot}52}{1-(0{\cdot}52)^2}=0{\cdot}7334$$

$$\therefore \qquad R_{1\cdot23}=+\,0{\cdot}8564$$

(since multiple correlation coefficient is non-negative).

Example 10·34. *In a trivariate distribution :*

$$\sigma_1=2,\,\sigma_2=\sigma_3=3,\; r_{12}=0{\cdot}7,\, r_{23}=r_{31}=0{\cdot}5.$$

Find (i) $r_{23\cdot1}$, *(ii)* $R_{1\cdot23}$, *(iii)* $b_{12\cdot3}$, $b_{13\cdot2}$, *and (iv)* $\sigma_{1\cdot23}$.

Solution. We have

$$(i)\quad r_{23\cdot1}=\frac{r_{23}-r_{21}r_{31}}{\sqrt{(1-r_{21}^2)\,(1-r_{31}^2)}}=\frac{0{\cdot}5-(0{\cdot}7)(0{\cdot}5)}{\sqrt{(1-0{\cdot}49)(1-0{\cdot}25)}}=0{\cdot}2425$$

$$(ii)\quad R_{1\cdot23}^2=\frac{r_{12}^2+r_{13}^2-2r_{12}\,r_{13}\,r_{23}}{1-r_{23}^2}$$

$$=\frac{0{\cdot}49+0{\cdot}25-2(0{\cdot}7)(0{\cdot}5)(0{\cdot}5)}{1-0{\cdot}25}=0{\cdot}52$$

$$\therefore \qquad R_{1\cdot23}=+\,0{\cdot}7211$$

$$(iii)\quad b_{12\cdot3}=r_{12\cdot3}\frac{\sigma_{1\cdot3}}{\sigma_{2\cdot3}} \quad\text{and}\quad b_{13\cdot2}=r_{13\cdot2}\frac{\sigma_{1\cdot2}}{\sigma_{3\cdot2}}$$

$$r_{12\cdot3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}=0{\cdot}6,\quad r_{13\cdot2}=\frac{r_{13}-r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}}=0{\cdot}2425$$

$$\sigma_{1\cdot3} = \sigma_1 \sqrt{(1-r_{13}^2)} = 2\sqrt{(1-0\cdot25)} = 1\cdot7320$$
$$\sigma_{2\cdot3} = \sigma_2 \sqrt{(1-r_{23}^2)} = 3\sqrt{(1-0\cdot25)} = 2\cdot5980$$
$$\sigma_{1\cdot2} = \sigma_1 \sqrt{(1-r_{12}^2)} = 2\sqrt{(1-0\cdot49)} = 1\cdot4282$$
$$\sigma_{3\cdot2} = \sigma_3 \sqrt{(1-r_{32}^2)} = 3\sqrt{(1-0\cdot25)} = 2\cdot5980$$

Hence $b_{12\cdot3} = 0\cdot4$ and $b_{13\cdot2} = 0\cdot1333$

(iv) $$\sigma_{1\cdot23} = \sigma_1 \sqrt{\frac{\omega}{\omega_{11}}}$$

where $$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} = 0\cdot36$$

and $$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - 0\cdot25 = 0\cdot75$$

$$\therefore \qquad \sigma_{1\cdot23} = 2 \times \sqrt{0\cdot48} = 2 \times 0\cdot6928 = 1\cdot3856$$

Example 10·35. *Find the regression equation of X_1 on X_2 and X_3 given the following results :—*

Trait	*Mean*	*Standard deviation*	r_{12}	r_{23}	r_{31}
X_1	*28·02*	*4·42*	*+ 0·80*	—	—
X_2	*4·91*	*1·10*	—	*−0·56*	—
X_3	*594*	*85*	—	—	*− 0·40*

where X_1 = *Seed per acre;* X_2 = *Rainfall in inches*

X_3 = *Accumulated temperature above 42°F.*

Solution. Regression equation of X_1 on X_2 and X_3 is given by

$$(X_1 - \bar{X}_1)\frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2)\frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3)\frac{\omega_{13}}{\sigma_3} = 0$$

where $$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - (-0\cdot56)^2 = 0\cdot686$$

$$\omega_{12} = -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13}\,r_{23} - r_{21} = -0\cdot576$$

$$\omega_{13} = r_{23}\,r_{12} - r_{13} = (-0\cdot56)(0\cdot80) - (-0\cdot40) = -0\cdot048$$

$\therefore$ Required equation of plane of regression of X_1 on X_2 and X_3 is given by

$$\frac{0\cdot686}{4\cdot42}(X_1 - 28\cdot02) + \frac{(-0\cdot576)}{1\cdot10}(X_2 - 4\cdot91) + \frac{(-0\cdot048)}{85\cdot00}(X_3 - 594) = 0$$

Example 10·36. *Five hundred students were examined in three subjects I, II and III, each subject carrying 100 marks. A student getting 120 or more but less than 150 marks was put in pass class. A student getting 150 or more but less than 180 marks was put in second class and a student getting 180 or more marks was put in the first class. The following marks were obtained :*

	I	II	II·
Mean :	*35·8*	*52·4*	*48·8*
S.D. :	*4·2*	*5·3*	*6·1*
Correlation :	$r_{12} = 0{\cdot}6,$	$r_{13} = 0{\cdot}7$	$r_{23} = 0{\cdot}8$

(i) Find the number of students in each of the three classes.

(ii) Find the total number of students with total marks lying between 120 and 190.

(iii) Find the probability that a student gets more that 240 marks.

(iv) What should be the correlation between marks in subjects I and II among students who scored equal marks in subject III ?

(v) If r_{23} was not known, obtain the limits within which it may lie from the values of r_{12} and r_{13} (ignoring sampling errors).

Solution. If Z denotes the total marks of the students in the three subjects and X_1, X_2, X_3 the total, marks of the students in subjects I, II and III respectively, then

$$Z = X_1 + X_2 + X_3$$

$$\therefore \quad E(Z) = E(X_1) + E(X_2) + E(X_3) = 35{\cdot}8 + 52{\cdot}4 + 48{\cdot}8 = 137$$

$$V(Z) = V(X_1) + V(X_2) + V(X_3) + 2[\text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_3) + \text{Cov}(X_3, X_1)]$$

$$= 17{\cdot}64 + 28{\cdot}09 + 37{\cdot}21 + 26{\cdot}712 + 35{\cdot}868 + 51{\cdot}728$$

$$= 197{\cdot}248 \qquad [\text{Using Cov}(X_i, X_j) = r_{ij}\sigma_i\sigma_j]$$

$$\Rightarrow \quad \sigma_Z^2 = 197{\cdot}248 \quad \text{or} \quad \sigma_Z = 14{\cdot}045$$

Now $$\xi = \frac{Z - E(Z)}{\sigma_Z} \sim N(0, 1)$$

Z	$\xi = \dfrac{Z - 137}{14{\cdot}045}$	$p = \int_{-\infty}^{\xi} p(\xi)d\xi$	*Class*	*Area under the curve in this class (A)*	*Frequency 500 × (A)*
120 –	1·21050	0·11314	120 – 150	0·70937	354·685
150	0·92567	0·82251	150 – 180	0·17639	88·195
180	3·06180	0·99890	180 –	0·00102	0·510
190	3·77400	0·99992	120 – 190	0·88678	443·390
240	7·33410	1·00000	240 –	0·00000	0·000

(*i*) The number of students in first, second and third class respectively are 355, 88 and 0 (approx.)

(*ii*) Total number of students with total marks between 120 and 190 is 443.

(*iii*) Probability that a student gets more than 240 marks is zero.

(*iv*) The correlation coefficient between marks in subjects I and II of the students who secured equal marks in subject III is $r_{12\cdot3}$ and is given by

$$r_{12\cdot3}=\frac{r_{12}-r_{13}\,r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}=\frac{0\cdot04}{\sqrt{(1-0\cdot49)(1-0\cdot64)}}=0\cdot0934$$

(v) We have

$$r_{12\cdot3}^2=\frac{(r_{12}-r_{13}\,r_{23})^2}{(1-r_{13}^2)(1-r_{23}^2)}\le 1$$

$$\therefore\qquad \frac{(0\cdot6-0\cdot7a)^2}{(1-0\cdot49)(1-a^2)}\le 1,\ \text{where } a=r_{23}.$$

$$\Rightarrow\qquad 0\cdot36+0\cdot49a^2-0\cdot84a\le 0\cdot51\,(1-a^2)$$

$$\Rightarrow\qquad a^2-0\cdot84a-0\cdot15\le 0$$

Thus 'a' lies between the roots of the equation :

$$a^2-0\cdot84a-0\cdot15=0,$$

which are 0·99 and – 0·15.

Hence r_{23} should lie between – 0·15 and 0·99.

Example 10·37. *Show that*

$$1-R_{1\cdot23}^2=(1-r_{12}^2)(1-r_{13\cdot2}^2)$$

Deduce that

(i) $R_{1\cdot23}\ge r_{12}$, (ii) $R_{1\cdot23}^2=r_{12}^2+r_{13}^2$, *if* $r_{23}=0$

(iii) $1-R_{1\cdot23}^2=\dfrac{(1-\rho)(1+2\rho)}{(1+\rho)}$, *provided all the coefficients of zero order are equal to* ρ.

(iv) *If* $R_{1\cdot23}=0$, X_1 *is uncorrelated with any of other variables, i.e.,* $r_{12}=r_{13}=0$. **[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]**

Solution. (i) Since $|\,r_{13\cdot2}\,|\le 1$, we have from (10·46c)

$$1-R_{1\cdot23}^2\le 1-r_{12}^2\quad\Rightarrow\quad R_{1\cdot23}\ge r_{12}$$

(ii) We have

$$r_{13\cdot2}=\frac{r_{13}-r_{12}\,r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}}=\frac{r_{13}}{\sqrt{1-r_{12}^2}}\qquad(\text{if } r_{23}=0)$$

$\therefore$ From (10·46c), we get

$$1-R_{1\cdot23}^2=(1-r_{12}^2)\left[1-\frac{r_{13}^2}{1-r_{12}^2}\right]=1-r_{12}^2-r_{13}^2$$

Hence $\qquad R_{1\cdot23}^2=r_{12}^2+r_{13}^2$, if $r_{23}=0$.

(iii) Here, we are given that $r_{12}=r_{13}=r_{23}=\rho$

$$\therefore\qquad r_{13\cdot2}=\frac{\rho-\rho^2}{\sqrt{(1-\rho^2)(1-\rho^2)}}=\frac{\rho(1-\rho)}{(1-\rho^2)}=\frac{\rho}{1+\rho}$$

Hence from (10·46c), we have

$$1-R_{1\cdot23}^2=(1-\rho^2)\left[1-\frac{\rho^2}{(1+\rho)^2}\right]=\frac{(1-\rho)(1+2\rho)}{(1+\rho)}$$

(iv) If $R_{1\cdot23}=0$, (10·46c) gives

$$1=(1-r_{12}^2)(1-r_{13\cdot2}^2)\qquad\ldots(*)$$

Since $0 \le r_{12}^2 \le 1$ and $0 \le r_{13.2}^2 \le 1$, (*) will hold if and only if

$$r_{12} = 0 \quad \text{and} \quad r_{13.2} = 0$$

Now $r_{13.2} = 0 \;\Rightarrow\; \dfrac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = 0$

$$\Rightarrow \quad \frac{r_{13}}{\sqrt{1 - r_{32}^2}} = 0 \qquad (\because r_{12} = 0)$$

$$\Rightarrow \quad r_{13} = 0$$

Thus if $R_{1.23} = 0$, then $r_{13} = r_{12} = 0$, *i.e.*, X_1 is uncorrelated with X_2 and X_3.

Example 10·38. *Show that the correlation coefficient between the residuals $X_{1.23}$ and $X_{2.13}$ is equal and opposite to that between $X_{1.3}$ and $X_{2.3}$.*

[*Poona Univ. B.Sc., 1991*]

Solution. The correlation coefficient between $X_{1.23}$ and $X_{2.13}$ is given by

$$\frac{\text{Cov}(X_{1.23}, X_{2.13})}{\sigma_{1.23}\,\sigma_{2.13}} = \frac{\Sigma X_{1.23} X_{2.13}}{N\,\sigma_{1.23}\,\sigma_{2.13}} = \frac{\frac{1}{N}\Sigma X_{2.13}(X_1 - b_{12.3} X_2 - b_{13.2} X_3)}{\sigma_{1.23}\,\sigma_{2.13}}$$

$$= -b_{12.3}\frac{\Sigma X_{2.13} X_2}{N\,\sigma_{1.23}\,\sigma_{2.13}} \qquad \text{(c.f. Property 1, § 10·13)}$$

$$= -b_{12.3}\frac{\Sigma X_{2.13}^2}{N\sigma_{1.23}\,\sigma_{2.13}} \qquad \text{(c.f. Property 2, § 10·13)}$$

$$= -b_{12.3}\frac{\sigma_{2.13}}{\sigma_{1.23}} = -b_{12.3}\frac{(\sigma_2\sqrt{\omega/\omega_{22}}\,)}{(\sigma_1\sqrt{\omega/\omega_{11}}\,)}$$

where $\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 \text{ and } \omega_{22} = \begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix} = 1 - r_{13}^2$$

$$\therefore\; r(X_{1.23}, X_{2.13}) = -b_{12.3}\frac{\sigma_2}{\sigma_1}\cdot\sqrt{\frac{1 - r_{23}^2}{1 - r_{13}^2}} = -b_{12.3}\frac{\sigma_{2.3}}{\sigma_{1.3}}$$

[since $\sigma_{2.3}^2 = \sigma_2^2(1 - r_{23}^2)$ and $\sigma_{1.3}^2 = \sigma_1^2(1 - r_{13}^2)$]

$$\therefore\; r(X_{1.23}, X_{2.13}) = -\frac{\text{Cov}(X_{1.3}, X_{2.3})}{\sigma_{2.3}^2}\cdot\frac{\sigma_{2.3}}{\sigma_{1.3}}$$

$$= -\frac{\text{Cov}(X_{1.3}, X_{2.3})}{\sigma_{2.3}\,\sigma_{1.3}} = -r(X_{1.3}, X_{2.3})$$

Hence the result.

Example 10·39. *Show that if $X_3 = aX_1 + bX_2$, the three partial correlations are numerically equal to unity, $r_{13.2}$ having the sign of a, $r_{23.1}$, the sign of b and $r_{12.3}$, the opposite sign of a/b.* **[*Kanpur Univ. M.Sc., 1992*]**

Solution. Here we may regard X_3 as dependent on X_1 and X_2 which may be taken as independent variables. Since X_1 and X_2 are independent, they are uncorrelated.

Thus $\quad r(X_1, X_2) = 0 \quad \Rightarrow \quad \text{Cov}(X_1, X_2) = 0$

$$V(X_3) = V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2ab\,\text{Cov}(X_1, X_2)$$
$$= a^2\sigma_1^2 + b^2\sigma_2^2,$$

where $\quad V(X_1) = \sigma_1^2,\ V(X_2) = \sigma_2^2.$

Also $\quad X_1X_3 = X_1(aX_1 + bX_2) = a\,X_1^2 + bX_1X_2$

Assuming that X_i's are measured from their means, on taking expectations of both sides, we get

$$\text{Cov}(X_1, X_3) = a\sigma_1^2 + b\,\text{Cov}(X_1, X_2) = a\sigma_1^2$$

$$\therefore \quad r_{13} = \frac{\text{Cov}(X_1, X_3)}{\sqrt{V(X_1)\,V(X_3)}} = \frac{a\,\sigma_1^2}{\sqrt{\sigma_1^2(a^2\sigma_1^2 + b^2\sigma_2^2)}} = \frac{a\sigma_1}{k},$$

where $\quad k^2 = a^2\sigma_1^2 + b^2\sigma_2^2.$

Similarly, we will get

$$r_{23} = \frac{\text{Cov}(X_2, X_3)}{\sqrt{V(X_2)\,V(X_3)}} = \frac{b\sigma_2}{k}$$

Hence

$$r_{13\cdot2} = \frac{r_{13} - r_{12}\,r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{a\sigma_1}{k}\,\frac{k}{\sqrt{k^2 - b^2\sigma_2^2}} = \frac{a\sigma_1}{\sqrt{a^2\sigma_1^2}} = \pm\frac{a\sigma_1}{a\sigma_1} = \pm 1$$

according as 'a' is positive or negative. Hence $r_{13\cdot2}$ has the same sign as 'a'.

Again

$$r_{23\cdot1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = \frac{b\sigma_2}{k}\,\frac{k}{\sqrt{k^2 - a^2\sigma_1^2}} = \frac{b\sigma_2}{\sqrt{b^2\sigma_2^2}} = \pm 1,$$

according as 'b' is positive or negative. Hence $r_{23\cdot1}$ has the same sign as 'b'.

Now

$$r_{12\cdot3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = -\frac{a\sigma_1}{k}\cdot\frac{b\sigma_2}{k}\cdot\frac{k^2}{\sqrt{(k^2 - a^2\sigma_1^2)(k^2 - b^2\sigma_2^2)}}$$
$$= -\frac{ab\sigma_1\sigma_2}{\sqrt{b^2\sigma_2^2 \times a^2\sigma_1^2}} = -\frac{ab}{\sqrt{a^2b^2}} = -\frac{a/b}{\sqrt{a^2/b^2}} = \frac{-(a/b)}{\pm(a/b)} = \mp 1,$$

according as (a/b) is positive or negative. Hence $r_{12\cdot3}$ has the sign opposite to that of (a/b).

Example 10·40. *If all the correlation coefficients of zero order in a set of p-variates are equal to ρ, show that*

(i) Every partial correlation of s' th order is $\dfrac{\rho}{1 + s\rho}$ $\qquad$ ...(*)

(ii) The coefficient of multiple correlation R of a variate with the other $(p - 1)$ variates is given by

$$1 - R^2 = (1 - \rho)\left[\frac{1 + (p - 1)\rho}{1 + (p - 2)\rho}\right]$$

[*Delhi Univ. M.Sc. (Maths), 1990*]

Solution. We are given that

$$r_{mn} = \rho, \ (m, n = 1, 2, \ldots, p \ ; m \neq n)$$

We have

$$r_{ij \cdot k} = \frac{r_{ij} - r_{ik}\, r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}, \quad (i, j, k = 1, 2, \ldots, p; \ i \neq j \neq k)$$

$$= \frac{\rho - \rho \cdot \rho}{\sqrt{(1 - \rho^2)(1 - \rho^2)}} = \frac{\rho}{1 + \rho} \qquad \ldots(**)$$

Thus every partial correlation coefficient of first order is $\rho/(1 + \rho)$.

$\Rightarrow$ (*) is true for $s = 1$.

The result will be established by the principle of mathematical induction. Let us suppose that every partial correlation coefficient of order s is given by $\rho/(1 + s\rho)$. Then the partial correlation coefficient of order $(s + 1)$ is given by

$$r_{ij \cdot km \ldots t} = \frac{r_{ij \cdot (s)} - r_{ik \cdot (s)} r_{jk \cdot (s)}}{\sqrt{(1 - r^2_{ik \cdot (s)})\,(1 - r^2_{jk \cdot (s)})}}$$

where $k, m, \ldots, t$ are $(s + 1)$ secondary subscripts and $r_{ij \cdot (s)}, r_{ik \cdot (s)}, r_{jk \cdot (s)}$, are partial correlation coefficients of order s. Thus

$$r_{ij \cdot km \ldots t} = \frac{\dfrac{\rho}{1 + s\rho} - \left(\dfrac{\rho}{1 + s\rho}\right)^2}{1 - \left(\dfrac{\rho}{1 + s\rho}\right)^2} = \frac{\dfrac{\rho}{1 + s\rho}\left(1 - \dfrac{\rho}{1 + s\rho}\right)}{\left(1 - \dfrac{\rho}{1 + s\rho}\right)\left(1 + \dfrac{\rho}{1 + s\rho}\right)} \ \frac{\rho}{1 + (s + 1)\rho}$$

Using (**) and (***), the required result follows by induction.

(*ii*) We have $1 - R^2 = \dfrac{\omega}{\omega_{11}}$

where R is the multiple correlation coefficient of a variable with other $(p - 1)$ variables and

$$\omega = \begin{vmatrix} 1 & \rho & \rho \ldots \rho \\ \rho & 1 & \rho \ldots \rho \\ \rho & \rho & 1 \ldots \rho \\ \vdots & \vdots & \vdots \quad \vdots \\ \rho & \rho & \rho \ldots 1 \end{vmatrix}, \text{ a determinant of order '}p\text{' and}$$

$$\omega_{11} = \begin{vmatrix} 1 & \rho & \rho \ldots \rho \\ \rho & 1 & \rho \ldots \rho \\ \rho & \rho & 1 \ldots \rho \\ \vdots & \vdots & \vdots \quad \vdots \\ \rho & \rho & \rho \ldots 1 \end{vmatrix} \quad \text{a determinant of order } (p - 1).$$

We have

$$\omega = [1 + (p-1)\rho] \begin{vmatrix} 1 & \rho & \rho & \rho & \dots & \rho \\ 1 & 1 & \rho & \rho & \dots & \rho \\ 1 & \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & \rho & \rho & \rho & \dots & 1 \end{vmatrix} \quad \text{(On adding } C_2, C_3 \dots, C_p \text{ to } C_1\text{).}$$

$$\Rightarrow \quad \omega = [1 + (p-1)\rho] \begin{vmatrix} 1 & \rho & \rho & \rho & \dots & \rho \\ 0 & (1-\rho) & 0 & 0 & \dots & 0 \\ 0 & 0 & (1-\rho) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \dots & (1-\rho) \end{vmatrix}$$

[On operating $R_i - R_1$, $(i = 2, 3, \dots p)$].

$$\therefore \qquad \omega = [1 + (p-1)\rho](1-\rho)^{p-1}$$

Similarly, we will have

$$\omega_{11} = [1 + (p-2)\rho](1-\rho)^{p-2}.$$

$$\therefore \qquad 1 - R^2 = \frac{\omega}{\omega_{11}} = (1-\rho)\left[\frac{1 + (p-1)\rho}{1 + (p-2)\rho}\right]$$

Example 10·41. *In a p-variate distribution, all the total (order zero) correlation coefficients are equal to $\rho_0 \neq 0$. Let ρ_k denote the partial correlation coefficient of order k and R_k be the multiple correlation coefficient of one variate on k other variates. Prove that*

(*i*) $\rho_0 \geq -\dfrac{1}{(p-1)}$, (*ii*) $\rho_k - \rho_{k-1} = -\rho_k \rho_{k-1}$, *and*

(*iii*) $R_k^2 = \dfrac{k\,\rho_0^2}{1 + (k-1)\rho_0}$. ***[Delhi Univ. M.Sc. (Stat.) 1987]***

Solution. (*i*) We have proved in Example 10·40, that

$$\rho_k = \frac{\rho_0}{1 + k\rho_0}$$

In the case of p-variate distribution, the partial correlation coefficient of the highest order is ρ_{p-2} and is given by

$$\rho_{p-2} = \frac{\rho_0}{1 + (p-2)\rho_0}$$

Since $\qquad |\rho_{p-2}| \leq 1 \;\Rightarrow\; -1 \leq \rho_{p-2} \leq 1,$

we have (on considering the lower limit)

$$-1 \leq \frac{\rho_0}{1 + (p-2)\rho_0} \quad \text{or} \quad -[1 + (p-2)\rho_0] \leq \rho_0$$

$$\Rightarrow \qquad \rho_0 \geq -\frac{1}{(p-1)}$$

(*ii*) L.H.S. $= \rho_k - \rho_{k-1} = \dfrac{\rho_0}{1+k\rho_0} - \dfrac{\rho_0}{1+(k-1)\rho_0}$

$$= \rho_0 \left[\frac{-\rho_0}{(1+k\rho_0)[1+(k-1)\rho_0]} \right]$$

$$= -\left(\frac{\rho_0}{1+k\rho_0}\right)\left(\frac{\rho_0}{1+(k-1)\rho_0}\right) = -\rho_k \rho_{k-1}$$

(*iii*) Taking $\rho = \rho_0$ and $k = p-1$ in part (*ii*) Example 10·40, we get

$$1 - R_k^2 = (1-\rho_0)\left[\frac{1+k\rho_0}{1+(k-1)\rho_0}\right]$$

$$\therefore \qquad R_k^2 = 1 - \frac{(1-\rho_0)(1+k\rho_0)}{1+(k-1)\rho_0} = \frac{k\,\rho_0^2}{1+(k-1)\rho_0} \quad \text{(On simplification).}$$

Example 10·42. *If r_{12} and r_{13} are given, show that r_{23} must lie in the range:* $r_{12}\, r_{13} \pm (1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2)^{1/2}$

If $r_{12} = k$ and $r_{13} = -k$, show that r_{23} will lie between -1 and $1 - 2k^2$.

[*Sardar Patel Univ. B.Sc. Oct., 1992; Madras Univ. B.Sc. (Stat. Main) 1991*]

Solution. We have

$$r_{12\cdot3}^2 = \left[\frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}\right]^2 \le 1$$

$$\therefore \qquad (r_{12} - r_{13}r_{23})^2 \le (1 - r_{13}^2)(1 - r_{23}^2)$$

$$\Rightarrow \qquad r_{12}^2 + r_{13}^2 r_{23}^2 - 2r_{12}\, r_{13}\, r_{23} \le 1 - r_{13}^2 - r_{23}^2 + r_{13}^2 r_{23}^2$$

$$\Rightarrow \qquad r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23} \le 1 \qquad \ldots(*)$$

This condition holds for consistent values of r_{12}, r_{13} and r_{23}. (*) may be rewritten as :

$$r_{23}^2 - (2r_{12}r_{13})r_{23} + (r_{12}^2 + r_{13}^2 - 1) \le 0.$$

Hence, for given values of r_{12} and r_{13}, r_{23} must lie between the roots of the quadratic (in r_{23}) equation

$$r_{23}^2 - (2r_{12}\, r_{13})\, r_{23} + (r_{12}^2 + r_{13}^2 - 1) = 0,$$

which are given by :

$$r_{23} = r_{12}r_{13} \pm \sqrt{r_{12}^2 r_{13}^2 - (r_{12}^2 + r_{13}^2 - 1)}$$

Hence

$$r_{12}\, r_{13} - \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2} \le r_{23} \le r_{12}\, r_{13} + \sqrt{(1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2)} \qquad \ldots(**)$$

In other words, r_{23} must lie in the range

$$r_{12}\, r_{13} \pm \sqrt{(1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2)}$$

In particular, if $r_{12} = k$ and $r_{13} = -k$, we get from (**)

$$-k^2 - \sqrt{(1 - k^2 - k^2 + k^4)} \le r_{23} \le -k^2 + \sqrt{(1 - k^2 - k^2 + k^4)}$$

$$\Rightarrow \quad -k^2-(1-k^2) \le r_{23} \le -k^2+(1-k^2)$$

$$\therefore \quad -1 \le r_{23} \le 1-2k^2$$

EXERCISE 10(g)

1\. (a) Explain partial correlation and multiple correlation.

(b) Explain the concepts of multiple and partial correlation coefficients.

Show that the multiple correlation coefficient $R_{1\cdot 23}$ is, in the usual notations given by : $R_{1\cdot 23}{}^2 = 1 - \dfrac{\omega}{\omega_{11}}$

2 (a) In the usual notations, prove that

$$R_{1\cdot 23}{}^2 = \frac{r_{12}{}^2 + r_{13}{}^2 - 2r_{12}r_{23}r_{31}}{1-r_{23}{}^2} \le r_{12}{}^2$$

(b) If $R_{1\cdot 23} = 1$, prove that $r_{2\cdot 13}$ is also equal to 1. If $R_{1\cdot 23} = 0$, does it necessarily mean that $R_{2\cdot 13}$ is also zero ?

3\. (a) Obtain an expression for the variance of the residual $X_{1\cdot 23}$ in terms of the correlations r_{12}, r_{23} and r_{31} and deduce that $R_{1(23)} \ge r_{12}$ and r_{13}.

(b) Show that the standard deviation of order p may be expressed in terms of standard deviation of order $(p-1)$ and a correlation coefficient of order $(p-1)$. Hence deduce that :

(i) $\sigma_1 \ge \sigma_{1\cdot 2} \ge \sigma_{1\cdot 23} \ge \ldots \ge \sigma_{1\cdot 23\ldots n}$

(ii) $1 - R^2_{1\cdot 23\ldots n} = (1-r^2_{12})(1-r^2_{13\cdot 2})\ldots(1-r^2_{1n\cdot 23\ldots(n-1)})$

[Delhi Univ. M.Sc. (Stat.) 1987]

4\. (a) In a p-variate distribution all the toal (zero order) correlation coefficients are equal to $\rho_0 \ne 0$. If ρ_k denotes the partial correlation coefficient of order k, find ρ_k. Hence deduce that :

(i) $\rho_k - \rho_{k-1} = -\rho_k \rho_{k-1}$

(ii) $\rho_0 \ge -1/(p-1)$. [Delhi Univ. M.Sc. (Stat.), 1989]

(b) Show that the multiple correlation coefficient $R_{1\cdot 23\ldots j}$ between X_1 and $(X_2, X_3, \ldots, X_j)$, $j = 2, 3, \ldots, p$ satisfies the inequalities :

$$R_{1\cdot 2} \le R_{1\cdot 23} \le \ldots \le R_{1\cdot 23\ldots p}$$

[Delhi Univ. M.Sc. (Maths.), 1989]

5\. (a) $X_0, X_1, \ldots, X_n$ are $(n+1)$ variates. Obtain a linear function of X_1, $X_2, \ldots, X_n$ which will have a maximum correlation with X_0. Show that the correlation R of X_0 with the linear function is given by

$$R = \left(1 - \frac{\omega}{\omega_{00}}\right)^{\frac{1}{2}}$$

where $$\omega = \begin{vmatrix} 1 & r_{01} & r_{02} \cdots\cdots & r_{0n} \\ r_{10} & 1 & r_{12} \cdots\cdots & r_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ r_{n0} & r_{n1} & r_{n2} \cdots\cdots & 1 \end{vmatrix}$$

and ω_{00} is the determinant obtained by deleting the first row and the first column of ω.

(b) With the usual notations, prove that

$$\sigma^2_{1\cdot 234\ldots n} = \frac{\omega}{\omega_{11}} \sigma_1^2 = \sigma_1^2 (1 - r_{12}^2)(1 - r_{13\cdot 2}^2)\ldots(1 - r^2_{1n\cdot 23\ldots n-1})$$

(c) For a trivariate distribution, prove that

$$r_{12\cdot 3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

6. (a) The simple correlation coefficients between temperature (X_1), corn yield (X_2) and rainfall (X_3) are, $r_{12} = 0{\cdot}59$, $r_{13} = 0{\cdot}46$ and $r_{23} = 0{\cdot}77$.

Calculate the partial correlation coefficients $r_{12\cdot 3}$, $r_{23\cdot 1}$ and $r_{31\cdot 2}$. Also calculate $R_{1\cdot 23}$.

(b) If $r_{12} = 0{\cdot}80$, $r_{13} = -0{\cdot}40$ and $r_{23} = -0{\cdot}56$, find the values of $r_{12\cdot 3}$, $r_{13\cdot 2}$ and $r_{23\cdot 1}$. Calculate further $R_{1(23)}$, $R_{2(13)}$ and $R_{3(12)}$.

7. (a) In certain investigation, the following values were obtained :

$$r_{12} = 0{\cdot}6,\ r_{13} = -0{\cdot}4 \text{ and } r_{23} = 0{\cdot}7$$

Are the values consistent ?

(b) Comment on the consistency of

$$r_{12} = \frac{3}{5},\ r_{23} = \frac{4}{5},\ r_{31} = -\frac{1}{2}.$$

(c) Suppose a computer has found, for a given set of values of X_1, X_2 and X_3,

$$r_{12} = 0{\cdot}91, \quad r_{13} = 0{\cdot}33 \quad \text{and} \quad r_{32} = 0{\cdot}81$$

Examine whether the computations may be said to be free from error.

8. (a) Show that if $r_{12} = r_{13} = 0$, then $R_{1(23)} = 0$. What is the significance of this result in regard to the multiple regression equation of X_1 on X_2 and X_3 ?

(b) For what value of $R_{1\cdot 23}$ will X_2 and X_3 be uncorrelated ?

(c) Given the data : $r_{12} = 0{\cdot}6$, $r_{13} = 0{\cdot}4$, find the value of r_{23} so that $R_{1\cdot 23}$, the multiple correlation coefficient of X_1 on X_2 and X_3 should be unity.

9. From the heights (X_1), weights (X_2) and ages (X_3) of a group of students the following standard deviations and correlation coefficients were obtained : $\sigma_1 = 2{\cdot}8$ inches, $\sigma_2 = 12$ lbs, and $\sigma_3 = 1{\cdot}5$ years, $r_{12} = 0{\cdot}75$, $r_{23} = 0{\cdot}54$, and $r_{31} = 0{\cdot}43$. Calculate (i) partial regression coefficients and (ii) partial correlation coefficients.

10. For a trivariate distribution :

$$\begin{array}{lll} \bar{X}_1 = 40 & \bar{X}_2 = 70 & \bar{X}_3 = 90 \\ \sigma_1 = 3 & \sigma_2 = 6 & \sigma_3 = 7 \\ r_{12} = 0{\cdot}4 & r_{23} = 0{\cdot}5 & r_{13} = 0{\cdot}6 \end{array}$$

Find

(*i*) $R_{1\cdot23}$, (*ii*) $r_{23\cdot1}$, (*iii*) the value of X_3 when $X_1 = 30$ and $X_2 = 45$.

11. (*a*) In a study of a random sample of 120 students, the following results are obtained :

$$\bar{X}_1 = 68, \quad \bar{X}_2 = 70, \quad \bar{X}_3 = 74$$
$$S_1^2 = 100, \quad S_2^2 = 25, \quad S_3^2 = 81,$$
$$r_{12} = 0{\cdot}60, \quad r_{13} = 0{\cdot}70, \quad r_{23} = 0{\cdot}65$$

[$S_i^2 = \text{Var}(X_i)$], where X_1, X_2, X_3 denote percentage of marks obtained by a student in I test, II test and the final examination respectively.

(*i*) Obtain the least square regression equation of X_3 on X_1 and X_2.

(*ii*) Compute $r_{12\cdot3}$ and $R_{3\cdot12}$.

(*iii*) Estimate the percentage marks of a student in the final examination if he gets 60% and 67% in I and II tests respectively.

(*b*) X_1 is the consumption of milk per head, X_2 the mean price of milk, and X_3, the per capita income. Time series of the three variables are rendered trend free and the standard deviations and correlation coefficients calculated :

$$s_1 = 7{\cdot}22, \quad s_2 = 5{\cdot}47, \quad s_3 = 6{\cdot}87$$
$$r_{12} = -0{\cdot}83, \quad r_{13} = 0{\cdot}92, \quad r_{23} = -0{\cdot}61$$

Calculate the regression equation of X_1 on X_2 and X_3 and interpret the regression as a demand equation.

12. (*a*) Five thousand candidates were examined in the subjects (*a*), (*b*), (*c*); each of these subjects carrying 100 marks. The following constants relate to these data :

	Subjects (*a*)	(*b*)	(*c*)
Mean	39·46	52·31	45·26
Standard deviation	6·2	9·4	8·7
	$r_{bc} = 0{\cdot}47$	$r_{ca} = 0{\cdot}38$	$r_{ab} = 0{\cdot}29$

Assuming normally correlated population, find the number of candidates who will pass if minimum pass marks are an aggregate of 150 marks for the three subjects together.

(*b*) Establish the equation of plane of regression for variates X_1, X_2, X_3 in the determinant form

$$\begin{vmatrix} X_1/\sigma_1 & X_2/\sigma_2 & X_3/\sigma_3 \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = 0$$

[*Delhi Univ. B.Sc. (Maths. Hons.), 1986*]

13. (*a*) Prove the identity

$$b_{12\cdot3}\, b_{23\cdot1}\, b_{31\cdot2} = r_{12\cdot3}\, r_{23\cdot1}\, r_{31\cdot2}$$ [*Gujarat Univ. B.Sc., 1992*]

(b) Prove that

$$R_{1\cdot 23}^2 = b_{12\cdot 3}\, r_{12} \frac{\sigma_2}{\sigma_1} + b_{13\cdot 2}\, r_{13} \frac{\sigma_3}{\sigma_1}$$

[*Sardar Patel Univ. B.Sc., 1991*]

14. (a) If $X_3 = aX_1 + bX_2$ for all sets of values of X_1, X_2, and X_3, find the value of $r_{23\cdot 1}$.

(b) If the relation $aX_1 + bX_2 + cX_3 = 0$ holds for all sets of values X_1, X_2 and X_3, what must be the partial correlation coefficients ?

15. (a) If $r_{12} = r_{23} = r_{31} = \rho \neq 1$, then

$$r_{12\cdot 3} = r_{23\cdot 1} = r_{31\cdot 2} = \frac{\rho}{1+\rho} \text{ and } R_{1(23)} = R_{2(13)} = R_{3(12)} = \frac{\rho\sqrt{2}}{\sqrt{(1+\rho)}}$$

(b) Y_1, Y_2, Y_3 are uncorrelated standard variates. $X_1 = Y_2 + Y_3$, $X_2 = Y_3 + Y_1$, and $X_3 = Y_1 + Y_2$. Find the multiple correlation coefficient between X_3 and (X_1 and X_2).

16. X, Y, Z are independent random variables with the same variance. If

$$X_1 = \frac{1}{\sqrt{2}}(X - Z),\; X_2 = \frac{1}{\sqrt{3}}(X + Y + Z),\; X_3 = \frac{1}{\sqrt{6}}(X + 2Y + Z),$$

show that X_1, X_2, X_3 have equal variances. Calculate $r_{12\cdot 3}$ and $R_{1(23)}$.

17. (a) If X_1, X_2 and X_3 are three variables measured from their respective means as origin and if e_1 is the expected value of X_1 for given values of X_2 and X_3 from the linear regression of X_1 on X_2 and X_3, prove that

$$\text{Cov}(X_1, e_1) = \text{Var}(e_1) = \text{Var}(X_1) - \text{Var}(X_1 - e_1)$$

(b) If $r_{12} = k$ and $r_{23} = -k$, show that r_{13} will lie between -1 and $1 - 2k^2$.

18. (a) For three variables X, Y and Z, prove that

$$r_{XY} + r_{YZ} + r_{ZX} \geq -\frac{3}{2} \qquad \ldots(*)$$

Hint. Let us transform X, Y, Z to their standard variables U, V and W (say), respectively, where

$$U = \frac{X - E(X)}{\sigma_X},\; V = \frac{Y - E(Y)}{\sigma_Y},\; W = \frac{Z - E(Z)}{\sigma_Z}$$

so that

$$\left.\begin{aligned} & E(U) = E(V) = E(W) = 0 \\ \text{and}\quad & \sigma_U^2 = \sigma_V^2 = \sigma_W^2 = 1 \;\Rightarrow\; E(U^2) = E(V^2) = E(W^2) = 1 \end{aligned}\right\} \ldots(**)$$

$$\left.\begin{aligned} & r_{UV} = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{E(UV) - E(U)E(V)}{\sigma_U \sigma_V} = E(UV) \\ \text{and}\quad & r_{UW} = E(UW)\;;\; r_{VW} = E(VW) \end{aligned}\right\} \ldots(***)$$

Since correlation coefficient is independent of change of origin and scale, proving (*) is equivalent to proving

$$r_{UV} + r_{VW} + r_{UW} \geq -3/2 \qquad \ldots(****)$$

To establish (****) let us consider the $E(U + V + W)^2$, which is always non-negative *i.e.*, $E(U + V + W)^2 \geq 0$, and use (**) and (***).

(*b*) X, Y, Z are three reduced (standard) variates and $E(YZ) = E(ZX) = -1/2$, find the limits between which the coefficient of correlation $r(X, Y)$ is necessarily placed.

Hint. Consider $E(X + Y + Z)^2 \geq 0 \quad \Rightarrow \quad r \geq -\frac{1}{2}$.

(*c*) If r_{12}, r_{23} and r_{31} are correlation coefficients of any three random variables X_1, X_2 and X_3 taken in pairs (X_1, X_2), (X_2, X_3) and (X_3, X_1) respectively, show that
$$1 + 2r_{12}\, r_{23}\, r_{31} \geq r_{12}^2 + r_{13}^2 + r_{23}^2$$

19. (*a*) If the relation $aX_1 + bX_2 + cX_3 = 0$, holds for all sets of values of X_1, X_2 and X_3, where X_1, X_2 and X_3 are three standardised variables, find the three total correlation coefficients r_{12}, r_{23} and r_{13} in terms of a, b and c. What are the values of partial correlation coefficients if a, b and c are positive ?

(*b*) Suppose X_1, X_2 and X_3 satisfy the relation $a_1X_1 + a_2X_2 + a_3X_3 = k$.

(*i*) Determine the three total correlation coefficients in terms of standard deviations and the constants a_1, a_2 and a_3.

(*ii*) State what the partial correlation coefficients would be.

20. (*a*) Show that the multiple correlation between Y and $X_1, X_2, \ldots, X_p$ is the maximum correlation between Y and any linear function of $X_1, X_2, \ldots, X_p$.

(*b*) Show that for p variates there are ${}^pC_2$ correlation coefficients of order zero and ${}^{p-2}C_s \cdot {}^pC_2$ of order s. Show further that there are ${}^pC_2 \cdot 2^{p-2}$ correlation coefficients altogether and ${}^pC_2 \cdot 2^{p-1}$ regression coefficients.

ADDITIONAL EXERCISES ON CHAPTER X

1. Find the correlation coefficient between

(*i*) $aX + b$ and Y, (*ii*) $lx + mY$ and $X + Y$, when correlation coefficient between X and Y is ρ.

2. If X_1 and X_2 are independent normal variates and U and V are defined by
$$U = X_1 \cos\alpha + X_2 \sin\alpha, \quad V = X_2 \cos\alpha - X_1 \sin\alpha,$$
show that the correlation coefficient ρ between U and V is given by
$$\rho^2 = 1 - \frac{4\sigma_1^2\sigma_2^2}{4\sigma_1^2\sigma_2^2 + (\sigma_1^2 - \sigma_2^2)\sin^2 2\alpha},$$
where σ_1^2 and σ_2^2 are variances of X_1 and X_2 respectively.

3. The variables X and Y are normally correlated, and ξ, η are defined by
$$\xi = X\cos\theta + Y\sin\theta, \quad \eta = Y\cos\theta - X\sin\theta$$
Obtain θ so that the distributions of ξ and η are independent.

4. A set of n observations of simultaneous values of X and Y are made by an observer and the standard deviations and product moment coefficient about the mean are found to be σ_X, σ_Y and ρ_{XY}. A second observer repeating the same observations made a constant error e in observing each X and a constant error E in observing each Y. The two sets of observations are combined into a single set and coefficient of correlation calculated from it. Show that its value is
$$\left(\rho_{XY} + \tfrac{1}{4}eE\right) + \sqrt{\left(\sigma_X^2 + \tfrac{1}{4}e^2\right)\left(\sigma_Y^2 + \tfrac{1}{4}E^2\right)}$$

Hint. here we have two sets of observations :

1st Set : (x, y_i), $i = 1, 2, \ldots, n$; Mean $= \bar{x}$, s.d. $= \sigma_x$.

Product moment coefficient $\rho_{xy} = r_{xy}\, \sigma_x \sigma_y$

2nd Set : $(x_i + e, y_i + E)$, $i = 1, 2, \ldots, n$

$$\text{Mean } (\bar{x}') = \frac{1}{N} \Sigma(x_i + e) = \bar{x} + e$$

$$\text{Variance} = \sigma_x^{2\prime} = \frac{1}{n} \Sigma\, [(x_i + e) - (\bar{x} + e)]^2 = \frac{1}{n} \Sigma\, (x_i - \bar{x})^2 = \sigma_x^2$$

$$\text{Mean } (\bar{y}') = \bar{y} + E,\ \sigma_y^{2\prime} = \sigma_y^2.$$

Product moment coefficient :

$$\rho_{xy}' = \frac{1}{n} \Sigma[(x_i + e) - (\bar{x} + e)][(y_i + E) - (\bar{y} + E)] = \rho_{xy}$$

To obtain the correlation coefficient for the combined set of $2n$ observations use Formula (10·5); Example 10·11(a) page 10·15.

5. Each of n independent trials can materialise in exactly one of the results $A_1, A_2, \ldots, A_k$. If the probability of A_i is p_i in every trial $\left(\sum\limits_{i=1} p_i = 1\right)$, find the probability of obtaining the frequencies $r_1, r_2, \ldots, r_k$ for $A_1, A_2, \ldots, A_k$ respectively in these trials. Also find $E(r_j)$, Var (r_j) and show that the correlation coefficient between r_i and r_j is independent of n.

6. In a sample of size n from a multinomial population $n_1, n_2, \ldots, n_k$ are of type $1, 2, \ldots, k$ with $\Sigma p_i = 1$, where p_i is the probability of type i $(i = 1, 2, \ldots, k)$. Show that the expected value of n_2 when n_1 is given is $(n - n_1)\, p_2(1 - p_1)$ and hence or otherwise show that the coefficient of correlation between n_i and n_j is

$$-\left[\frac{p_i p_j}{(1 - p_i)\,(1 - p_j)}\right]^{\frac{1}{2}}$$

7. A ball is drawn at random from an urn containing 3 white balls numbered 0, 1, 2 ; 2 red balls numbered 0, 1 and 1 black ball numbered 0. If the colours white, red and black are again numbered 0. 1 and 2 respectively, show that the correlation coefficient between the variables : X, the colour number and Y, the number of the ball is $-\frac{1}{2}$.

8. If X_1 and X_2 are two independent normal variates with a common mean zero and variances σ_1^2 and σ_2^2 respectively, show that the variates defined by

$$U_1 = X_1 + X_2 \quad \text{and} \quad U_2 = -\frac{\sigma_2}{\sigma_1} X_1 + \frac{\sigma_1}{\sigma_2} X_2$$

are independent and that each is normally distributed with mean zero and common variance $(\sigma_1^2 + \sigma_2^2)$.

9. If X_1, X_2 and X_3 are uncorrelated variables with equal mean M and variances V_1^2, V_2^2 and V_3^2 respectively, prove that correlation coefficient ρ between $Z_1 = \dfrac{X_1}{X_3}$ and $Z_2 = \dfrac{X_2}{X_3}$ is given by

$$\rho = \frac{V_3^2}{\sqrt{[(V_1^2 + V_3^2)(V_2^2 + V_3^2)}}$$

Hint. Neglecting the cubes and higher powers of $\dfrac{x_i}{M}$, x_i being the deviation of X_i from M and letting the means and s.d.'s of Z_1 and Z_2 to be I_1, I_2 and s_1, s_2 respectively, we get

$$I_1 = \frac{1}{N}\sum \frac{X_1}{X_3} = \frac{1}{N}\sum_i (x_{1i} + M)(x_{3i} + M)^{-1}$$

$$= \frac{1}{N}\sum \left(1 + \frac{x_{1i}}{M}\right)\left(1 + \frac{x_{3i}}{M}\right)^{-1}$$

$$= \frac{1}{N}\sum \left[\left(1 - \frac{x_{3i}}{M} + \frac{x_{3i}^2}{M^2} - \dots\right) + \frac{x_{1i}}{M} - \frac{x_{1i}x_{3i}}{M^2} + \dots\right]$$

$$= 1 + \frac{V_3^2}{M^2}$$

Similarly $I_2 = 1 + \dfrac{V_3^2}{M^2}$

$\therefore \quad I_1 = I_2$

Now $\quad s_1^2 = \dfrac{1}{N}\sum\left(\dfrac{X_1}{X_3}\right)^2 - I_1^2$

or $\quad s_1^2 + I_1^2 = 1 + \dfrac{3V_3^2}{M^2} + \dfrac{V_1^2}{M^2}$, and so we have $s_1^2 = \dfrac{V_3^2}{M^2} + \dfrac{V_1^2}{M^2}$.

Similarly $s_2^2 = \dfrac{V_3^2}{M^2} + \dfrac{V_2^2}{M^2}$

Now $N\rho s_1 s_2 = \sum\left(\dfrac{X_1}{X_3} - I_1\right)\left(\dfrac{X_2}{X_3} - I_2\right) = \dfrac{V_3^2}{M^2}$ (On simplification)

Hence $\quad \rho = \dfrac{N\rho\, s_1 s_2}{s_1 s_2} = \dfrac{V_3^2}{\sqrt{(V_3^2 + V_1^2)}\,\sqrt{(V_3^2 + V_2^2)}}$

10. (*Weldon's Dice Problem*). n white dice and m red dice are shaken together and thrown on a table. The sum of the dots on the upper faces are noted. The red dice are then picked up and thrown again among the white dice left on the table. The sum of the dice on the upper faces is again noted. What is the correlation coefficient between the first and the second sums ?

Ans. $n/(n + m)$

11. Random variables X and Y have zero means and non-zero variances σ_X^2 and σ_Y^2. If $Z = Y - X$, then find σ_Z^2 and the correlation coefficient $\rho(X, Z)$ of X and Z in terms of σ_X, σ_Y and the correlation coefficient $r(X, Y)$ of X and Y.

For certain data $Y = 1{\cdot}2X$ and $X = 0{\cdot}6Y$, are the regression lines. Compute $r(X, Y)$ and σ_X/σ_Y. Also compute $\rho(X, Z)$, if $Z = Y - X$.

[Calcutta Univ. B.Sc. (Maths. Hons.), 1984]

12. An item (say, a pen) from a production line can be acceptable, repairable or useless. Suppose a production is stable and let p, q, r $(p + q + r = 1)$, denote the probabilities for three possible conditions of an item. If the items are put into lots of 100 :

(i) Derive an expression for the probability function of (X, Y) where X and Y are the number of items in the lots that are respectively in the first two conditions.

(ii) Derive the moment generating function of X and Y.

(iii) Find the marginal distribution X.

(iv) Find the conditional distribution of Y given $X = 90$.

(v) Obtain the regression function of Y on X.

[Delhi Univ. M.A. (Eco.), 1985]

13. If the regression of X_1 on $X_2, \ldots, X_p$ is given by :

$$E(X_1 \mid X_2, \ldots, X_p) = \alpha + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

and

$$\begin{vmatrix} \sigma_{22} & \sigma_{23} \cdots & \sigma_{2p} \\ \sigma_{32} & \sigma_{33} \cdots & \sigma_{3p} \\ \vdots & \vdots & \vdots \\ \sigma_{p2} & \sigma_{p3} \cdots & \sigma_{pp} \end{vmatrix} > 0, \quad \begin{pmatrix} \sigma_{ii} = \text{variances} \\ \sigma_{ij} = \text{covariances} \end{pmatrix}$$

then the constants $\alpha, \beta_2, \ldots, \beta_p$ are given by

$$\alpha = \mu_1 + \frac{R_{12}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_2} \cdot \mu_2 + \frac{R_{13}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_3} \cdot \mu_3 + \ldots + \frac{R_{1p}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_p} \cdot \mu_p$$

and

$$\beta_j = -\frac{R_{1j}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_j}, \ (j = 1, 2, \ldots, p)$$

where R_{ij} is the cofactor of ρ_{ij} in the determinant (R) of the correlation matrix

$$R = \begin{vmatrix} \rho_{11} & \rho_{12} \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} \cdots & \rho_{pp} \end{vmatrix}$$

[Delhi Univ. M.Sc. (Stat.), 1989]

14. Let X_1 and X_2 be random variables with means 0 and variances 1 and correlation coefficient ρ. Show that :

$$E[\max(X_1^2, X_2^2)] \le 1 + \sqrt{1 - \rho^2}$$

Using the above inequality, show that for random variables X_1 and X_2 with means μ_1 and μ_2, variances σ_1^2 and σ_2^2 and correlation coefficient ρ and for any $k > 0$,

$$P\left[|X_1-\mu_1|\ge k\sigma_1 \text{ or } |X_2-\mu_2|\ge k\sigma_2\right]\le \frac{1}{k^2}\left[1+\sqrt{1-\rho^2}\right]$$

15. Let the maximum correlation between X_0 and any linear function of $X_1, X_2, \ldots, X_n$ be R and if $r_{01}=r_{02}=\ldots=r_{0n}=r$ and all other correlation coefficients are equal to s, then show that :

$$R=r\left[\frac{n}{1+(n-1)s}\right]^{1/2}$$

16. If $f=f(x, y)$ is the p.d.f. of $BVN\ (0, 0, 1, 1, \rho)$ distribution, verify that :

$$\frac{\partial f}{\partial \rho}=\frac{\partial^2 f}{\partial x\partial y}$$

Further, if two new random variables U and V are defined by the relation

$$U=P(Z\le x) \text{ and } V=P(Z\le y) \text{ where } Z\sim N(0, 1),$$

prove the marginal distributions of both U and V are uniform in the interval $\left(-\frac{1}{2}, \frac{1}{2}\right)$ and their common variance is $\frac{1}{12}$.

Hence prove that $R=$ Corr. (U, V), satisfies the relation : $\rho=2\sin(\pi R/6)$.

[Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1988]

17. If $(X, Y)\sim BVN\ (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, then prove that $a+bX+cY$, $(b\ne 0, c\ne 0)$ is distributed as $N(a+b\mu_x+c\mu_y, b^2\sigma_x^2+c^2\sigma_y^2+2bc\rho\sigma_x\sigma_y)$.

[Delhi Univ. M.Sc. (Stat.), 1989]

18. Let X_1, X_2, X_3 be a random sample of size $n=3$ from $N(0, 1)$ distribution.

(*a*) Show that $Y_1=X_1+\delta X_3$, $Y_2=X_2+\delta X_3$ has a bivariate normal distribution.

(*b*) Find the value of δ so that $\rho(Y_1, Y_2)=\frac{1}{2}$.

(*c*) What additional transformation involving Y_1 and Y_2 would produce a bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and the same correlation coefficient ρ ?

Ans. (*b*) -1 or 1. (*c*) $Z_1=\sigma_1 Y_1+\mu_1$, $Z_2=\sigma_2 Y_2+\mu_2$.

19. If $(X, Y)\sim BVN\ (0, 0, 1, 1, \rho)$, prove that :

$$E[\max(X, Y)]=[(1-\rho)/\pi]^{1/2} \text{ and } E[\min(X, Y)]=-[(1-\rho)/\pi]^{1/2}$$

20. If $(X, Y)\sim BVN\ (0, 0, \sigma_1^2, \sigma_2^2, \rho)$, show that rth cumulant of XY is given by :

$$\kappa_r=\frac{1}{2}(r-1)!\ \sigma_1^r\sigma_2^r\left[(\rho+1)^r+(\rho-1)^r\right].$$

Deduce that $E(X^2Y^2)=\sigma_1^2\sigma_2^2(1+2\rho^2)$.

21. Let f and g be the p.d.f.'s of X and Y with corresponding distribution functions F and G. Also let

$$h(x, y)=f(x)\,g(y)\left[1+\alpha\,(2f(x)-1)\,(2G(x)-1)\right];\ |\alpha|\le 1,$$

Show that $h(x, y)$ is a joint p.d.f. with marginal p.d.f.'s f and g. Further, let f and g be $N(0, 1)$ p.d.f.'s. Show that $Z=X+Y$, is not normally distributed, except in the trivial case $\alpha=0$.

Hint. Find $M_Z(t) = E(e^{tZ})$ and use Cov $(X, Y) = \alpha/\pi$.

22. State p.d.f. of bivariate normal distribution. Let X and Y have joint p.d.f. of the form :

$$f(x, y) = ke^{-\frac{1}{2}[a_{11}(x-b_1)^2 + 2a_{12}(x-b_1)(y-b_2) + a_{22}(y-b_2)^2]}; \quad -\infty \le (x, y) <$$

Find (*i*) k, (*ii*) the correlation coefficient between X and Y.

23. Write down, but do not derive, the moment generating function for a pair of random variables which have a bivariate normal distribution with both means equal to zero.

The independent random variables X, Y, Z, are each normally distributed with mean 0 and variance 1. If $U = X + Y + Z$ and $V = X - Y + 2Z$, show that U and V have bivariate normal distribution. Find the correlation of U with V and the expectation of U when V is equal to 1.

24. Let X_1 and X_2 have a joint m.g.f.

$$M(t_1, t_2) = \left[a(e^{t_1 + t_2} + 1) + b(e^{t_1} + e^{t_2})\right]^2$$

in which a and b are positive constants such that $2a + 2b = 1$.

Find $E(X_1)$, $E(X_2)$, Var (X_1), Var (X_2), Cov (X_1, X_2).

Ans. Means = 1, Variances = $\frac{1}{2}$; Covariance = $2a - \frac{1}{2}$.

25. X_1, X_2, X_3 have joint distribution as a multinomial distribution with parameters N, p_1, p_2, p_3. If r_{ij} is the correlation coefficient between X_i and X_j, find the expression for r_{12}, r_{23} and r_{31} and hence deduce the expression for the partial correlation coefficient $r_{1\cdot 23}$.

26. (*i*) If all the inter-correlations between $(p + 1)$ variates $X_0, X_1, X_2, \ldots, X_p$ are equal to r, show that each of the partial correlation co-efficients of order $p - 1$ is equal to $r/[1 + (p - 1)r]$ and that the multiple correlation of X_0 on $X_1, X_2, \ldots, X_p$ is given by

$$1 - R^2_{0(12\ldots p)} = \frac{(1-r)(1-pr)}{1 + (p-1)r}$$

(*ii*) $\quad r_{12} = (r_{12\cdot 3} - r_{13\cdot 2}\, r_{23\cdot 1}) / [(1 - r^2_{13\cdot 2})^{1/2}(1 - r^2_{23\cdot 1})^{1/2}]$

27. If R denotes the multiple correlation co-efficient of X_1 on $X_2, X_3, \ldots, X_p$ in p-variate distribution, prove that

(*i*) $R^2 \ge R_0^2$, where R_0 is the correlation of X_1 with any arbitrary linear function of $X_2, X_3, \ldots, X_p$.

(*ii*) $R^2 \ge R_1^2$, where R_1 is the multiple correlation coefficient of X_1 with $X_2, X_3, \ldots, X_k$, $k < p$

(*iii*) $\quad 1 - R^2 = \prod_{j=2}^{p} (1 - r^2_{1j\cdot 23\ldots(j-1)})$

Theory of Attributes

11·1. Introduction. Literally, an attribute means a *quality* or *characteristic*. Theory of attributes deals with qualitative characteristics which are not amenable to quantitative measurements and hence need slightly different statistical treatment from that of the variables. Examples of attributes are drinking, smoking, blindness, health, honesty, etc. An attribute may be marked by its presence (possession) or absence (dispossession) in a member of given population. It may be pointed out that the method of statistical analysis applicable to the study of variables can also be used to a great extent in the theory of attributes and *vice-versa*. For example, the presence or absence of an attribute may be regarded as changes in the values of a variable which can possess only two values *viz*. 0 and 1. For the sake of clarity and simplicity, the theory of attributes has been developed independently.

11·2. Notations. Suppose the population is divided into two classes, according to the presence or absence of a single attribute. The *positive class*, which denotes the presence of the attribute is generally written in capital Roman letters such as A, B, C, D etc. and the *negative class*, denoting the absence of the attribute is written in corresponding small Greek letters such as $\alpha, \beta, \gamma, \delta$, etc. For example if A represents the attribute sickness and B represents blindness, than α and β represent the attributes non-sickness (health) and sight respectively. The two classes *viz*., A (possession of the attribute) and α (dispossession of the attribute) are said to be *complementary classes* and the attribute α used in the sense of not-A is often called the complementary attribute of A. Similarly β, γ, δ are the complementary attributes of B, C, D respectively.

The combinations of attributes are denoted by grouping together the letters concerned *e.g.* AB is the combination of the attributes A and B. Thus for the attributes A (sickness) and B (smoking), AB would mean the simultaneous possession of sickness and smoking. Similarly $A\beta$ will represent sickness and non-smoking, αB non-sickness (health) and smoking, and $\alpha\beta$ non-sickness and non-smoking.

If a third attribute be included to represent, say male, then ABC will stand for sick males who are smokers. Similar interpretations can be given to $AB\gamma$, $A\beta C$, $A\beta\gamma$ etc.

11·3. Dichotomy. If the universe (population) is divided into two sub-classes or complementary classes and no more, with respect to each of the attributes A, B, C etc., the division or classification is said to be '*dichotomous classification*". The classification is termed *manifold* if each class is further sub-divided.

11·4. Classes and Class Frequencies. Different attributes in themselves are called different classes and the number of observations assigned to them are called *class frequencies* which are denoted by bracketing the class-symbols. Thus (A) stands for the frequency of A and (AB) for the number of objects possessing the attribute AB.

Remark. Class frequencies of the type (A), (AB), (ABC) etc. are known as *positive frequencies;* (α), $(\alpha\beta)$, $(\alpha\beta\gamma)$ etc. are known as *negative frequencies* and (αB), $(A\beta C)$, $(\alpha\beta C)$ etc. are called the *contrary frequencies.*

11·4·1. Order of Classes and Class frequencies. A class represented by n attributes is called a class of nth order and the corresponding frequency as the frequency of the nth order. Thus (A) is a class frequency of order 1; (AB), (AC), $(\beta\gamma)$ etc. are class frequencies of second order; (ABC), $(A\beta\gamma)$ $(\alpha\beta C)$ etc. are frequencies of third order and so on. N, the total number of members of the population, without any specification of attributes, is reckoned as a frequency of zero-order.

Thus in a dichotomous classification with respect to n attributes, the number of class frequencies of order 'r' is $\binom{n}{r} . 2^r$, since r attributes out of n can be selected in $\binom{n}{r}$ ways and each of the r attributes contributes two symbols, one representing the positive part (*e.g.*, A) and the other the negative part (*e.g.*, α). Thus the total number of class frequencies of all orders, for n attributes is :

$$\sum_{r=0}^{n} \binom{n}{r} 2^r = 1 + \binom{n}{1} 2 + \binom{n}{2} 2^2 + \ldots + \binom{n}{n} 2^n = (1+2)^n = 3^n \qquad \ldots(11·1)$$

Remarks 1. In particular, for n attributes, the total number of class frequencies of different orders are given as follows :

Order	0	1	2	...	r	...	n
No. of frequencies	1	$2n$	$\binom{n}{2} 2^2$	...	$\binom{n}{r} 2^r$	...	2^n

2. Since in the case of n attributes, the positive class frequency of order r has $\binom{n}{r}$ elements, their total number is :

$$\sum_{r=0}^{n} \binom{n}{r} = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n} = (1+1)^n = 2^n$$

3. In case of 3 attributes A, B and C, the total number of class frequencies is $3^3 = 27$, as given below :

$$
\left.\begin{array}{cllll}
\textbf{Order} & & & \textbf{Frequencies} & \\
0 & & & N & \\
1 & \left\{\begin{array}{l}(A)\\(\alpha)\end{array}\right. & \begin{array}{l}(B)\\(\beta)\end{array} & \begin{array}{l}(C)\\(\gamma)\end{array} & \\
2 & \left\{\begin{array}{l}(AB)\\(AC)\\(BC)\end{array}\right. & \begin{array}{l}(A\beta)\\(A\gamma)\\(B\gamma)\end{array} & \begin{array}{l}(\alpha B)\\(\alpha C)\\(\beta C)\end{array} & \begin{array}{l}(\alpha\beta)\\(\alpha\gamma)\\(\beta\gamma)\end{array} \\
3 & \left\{\begin{array}{l}(ABC)\\(\alpha BC)\end{array}\right. & \begin{array}{l}(AB\gamma)\\(\alpha B\gamma)\end{array} & \begin{array}{l}(A\beta C)\\(\alpha\beta C)\end{array} & \begin{array}{l}(A\beta\gamma)\\(\alpha\beta\gamma)\end{array}
\end{array}\right\} \qquad ...(11\cdot2)
$$

11·4·2. Relation Between Class Frequencies. All the class frequencies of various orders are not independent of each other and any class frequency can always be expressed in terms of class frequencies of hi- . -der. Thus

$$N = (A) + (\alpha) = (B) + (\beta) = (C) + (\gamma), \text{ etc.}$$

Also, since each of these A's or α's can either be B's or β's, we have

$$(A) = (AB) + (A\beta) \quad \text{and} \quad (\alpha) = (\alpha B) + (\alpha\beta)$$

Similarly
$$(B) = (AB) + (\alpha B) \quad \text{and} \quad (\beta) = (A\beta) + (\alpha\beta)$$
$$(AB) = (ABC) + (AB\gamma), \qquad (A\beta) = (A\beta C) + (A\beta\gamma)$$
$$(\alpha B) = (\alpha BC) + (\alpha B\gamma), \qquad (\alpha\beta) = (\alpha\beta C) + (\alpha\beta\gamma)$$

and so on. Thus

$$(A) = (AB) + (A\beta) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$$
$$(\beta) = (A\beta) + (\alpha\beta) = (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma), \text{ etc.}$$

The classes of highest order are called the *ultimate classes* and their frequencies, the *ultimate class frequencies*. Thus in case of n attributes, the ultimate class frequencies will be the frequencies of nth order. For example, the class frequencies (ABC), $(AB\gamma)$, $(A\beta C)$, $(A\beta\gamma)$, (αBC), $(\alpha B\gamma)$, $(\alpha\beta C)$, $(a\beta\gamma)$ are the ultimate frequencies for three attributes A, B and C.

Remarks 1. In case of n attributes, the ultimate class frequencies each contain n symbols and since each symbol may be written in two ways, *viz.*, positive part and negative part, *e.g.*, A or α, B or β, etc., the total number of ultimate class frequencies is 2^n.

2. By expressing any class frequency in terms of the class frequency of higher order, we can express it ultimately as the sum of some of the 2^n ultimate class frequencies.

3. The total number of ultimate class frequencies specify the data completely.

4. The set of ultimate class frequencies is not the only set which specify the data completely. In fact any set of class frequencies which are (*i*) 2^n in number and (*ii*) which are algebraically independent of each other, will specify the data completely. Such a set is called the *Fundamental Set.* For example, the positive class frequencies form such a set. Thus for $n = 2$, the set of positive class frequencies $2^2 = 4$ (c.f. 11·2), is N, (A), (B), (AB). If we are given these

frequencies, then it is obvious from the table that the remaining frequencies, *viz.*, $(A\beta)$, $(\alpha\beta)$ (αB),(α) and (β) can be obtained by subtraction, *e.g.*, given :

	A	α	Total
B	(AB)	–	(B)
β	–	–	(β)
Total	(A)	(α)	N

$$(\alpha) = N - (A),\ (\beta) = N - (B)$$
$$(A\beta) = (A) - (AB),\ (\alpha B) = (B) - (AB)$$
$$(\alpha\beta) = (\alpha) - (\alpha B) = N - (A) - (B) + (AB)$$

11.5. Class Symbols as Operators. Let us write symbolically

$$A.N = (A) \quad \ldots(11\cdot3)$$

which means that the operation of dichotomising N according to A gives the class frequency equal to (A). Similarly, we write

$$\alpha.N = (\alpha)$$

Adding, we get

$$A.N + \alpha.N = (A) + (\alpha)$$
$$\Rightarrow \quad (A + \alpha).\ N = (A) + (\alpha)$$
$$\Rightarrow \quad (A + \alpha).\ N = N$$
$$\Rightarrow \quad A + \alpha = 1$$

Thus in symbolic expression we can replace A by $(1 - \alpha)$ and α by $(1 - A)$. Similarly, B can be replaced by $(1 - \beta)$ and β by $(1 - B)$, and so on.

Dichotomising (B) according to A, let us write

$$A.\ (B) = (AB)$$

Similarly, $$B.\ (A) = (BA)$$

$$A.\ (B) = B.\ (A) = (AB) = AB.N,$$

which amounts to dichotomising N according to AB.

For example :

$$(\alpha\beta) = \alpha\beta.\ N = (1 - A)\ (1 - B).\ N = N - A.\ N - B.\ N + AB.\ N$$
$$= N - [\ (A) + (B)] + (AB)$$
$$(\alpha\beta\gamma) = \alpha\beta\gamma.\ N = (1 - A)\ (1 - B)\ (1 - C).\ N$$
$$= N - A.N - B.N - C.N + AB.N + AC.N + BC.N - ABC.N$$
$$= N - [(A) + (B) + (C)] + [(AB) + (AC) + (BC)] - (ABC)$$
$$(AB\gamma) = AB\gamma.\ N = AB(1 - C).\ N = AB.\ N - ABC.\ N$$
$$= (AB) - (ABC)$$
$$(\alpha\beta C) = (1 - A)\ (1 - B)\ C.\ N = (C - AC - BC + ABC).\ N$$
$$= (C) - (AC) - (BC) + (ABC)$$

and so on.

Example 11·1. *An investigation of 23,713 households was made in an urban and rural mixed locality. Of these 1,618 were farmers, 2,015 well-to-do and 770 families were having at least one graduate. Of these graduate families 335 were those of farmers and 428 were well-to-do, also 587 well-to-do families were those of farmers and out of them only 156 were having at least one of their family member as graduate. Obtain all the ultimate class frequencies.*

Solution. Let the attribute 'farming' be denoted by A, the attribute 'well-to-do' by B and 'having at least one graduate' by C. Then in the usual notations, we are given

$$N = 23713,\ (A) = 1618,\ (B) = 2015,\ (C) = 770,\ (AB) = 587,$$
$$(BC) = 428,\quad (AC) = 335\quad \text{and}\quad (ABC) = 156.$$

For three attributes A, B, C, the number of ultimate class frequencies is $2^3 = 8$, one of them being $(ABC) = 156$. The remaining frequencies are obtained below :

$$\begin{aligned}
(AB\gamma) &= (AB) - (ABC) = 587 - 156 = 431\\
(A\beta C) &= (AC) - (ABC) = 335 - 156 = 179\\
(A\beta\gamma) &= (A) - (AB) - (AC) + (ABC)\\
&= 1618 - 587 - 335 + 156 = 852\\
(\alpha BC) &= (BC) - (ABC) = 428 - 156 = 272\\
(\alpha B\gamma) &= (B) - (AB) - (BC) + (ABC)\\
&= 2015 - 587 - 428 + 156 = 1156\\
(\alpha\beta C) &= (C) - (AC) - (BC) + (ABC) = 770 - 335 - 428 + 156 = 163\\
(\alpha\beta\gamma) &= N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC)\\
&= 23713 - 1618 - 2015 - 770 + 587 + 335 + 428 - 156 = 20504
\end{aligned}$$

Example 11·2. (a) *Given the following ultimate class frequencies, find the frequencies of positive class,*

$$(ABC) = 149,\quad (AB\gamma) = 738,\quad (A\beta C) = 225,\quad (A\beta\gamma) = 1{,}196$$
$$(\alpha BC) = 204,\quad (\alpha B\gamma) = 1{,}762,\quad (\alpha\beta C) = 171\quad \text{and}\quad (\alpha\beta\gamma) = 21{,}842$$

(b) *Find the remaining class frequencies, given the following data :*

$$N = 23{,}713,\quad (A) = 1618,\quad (B) = 2015,\quad (C) = 770$$
$$(AB) = 587,\quad (AC) = 428,\quad (BC) = 335,\quad (ABC) = 156$$

Solution. (*a*)
$$\begin{aligned}
(A) &= (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) = 2{,}308\\
(B) &= (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma) = 2{,}853\\
(C) &= (ABC) + (A\beta C) + (\alpha BC) + (\alpha\beta C) = 749\\
(AB) &= (ABC) + (AB\gamma) = 887\\
(AC) &= (ABC) + (A\beta C) = 374\\
(BC) &= (ABC) + (\alpha BC) = 353
\end{aligned}$$

and
$$N = [(ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)] = 26{,}287$$

(*b*) For three attributes, there are $3^3 = 27$, class frequencies in all. Thus we have to determine the remaining 19 class frequencies :

Order 1 :

$$(\alpha) = N - (A) = 22{,}095\ ;\ (\beta) = N - (B) = 21{,}698$$

$(\gamma) = N - (C) = 22{,}943$

Order 2 :

$(A\beta) = (A) - (AB) = 1{,}031$
$(\alpha B) = (B) - (AB) = 1{,}428$
$(\alpha\beta) = (\alpha) - (\alpha B) = 20{,}667$
$(A\gamma) = (A) - (AC) = 1{,}190$
$(\alpha C) = (C) - (AC) = 342$
$(\alpha\gamma) = (\alpha) - (\alpha C) = 21{,}753$
$(B\gamma) = (B) - (BC) = 1{,}680$
$(\beta C) = (C) - (BC) = 435$
$(\beta\gamma) = (\beta) - (\beta C) = 21{,}263$

Order 3 :

$(AB\gamma) = (AB) - (ABC) = 431$
$(A\beta C) = (AC) - (ABC) = 272$
$(A\beta\gamma) = (A\beta) - (A\beta C) = 759$
$(\alpha BC) = (BC) - (ABC) = 179$
$(\alpha B\gamma) = (\alpha B) - (\alpha BC) = 1249$
$(\alpha\beta C) = (\beta C) - (A\beta C) = 163$
$(\alpha\beta\gamma) = (\alpha\beta) - (\alpha\beta C) = 20{,}504$

Example 11·3. *Show that for n attributes $A_1, A_2, A_3, \ldots, A_n$*

$$(A_1 A_2 A_3 \ldots A_n) \geq (A_1) + (A_2) + (A_3) + \ldots + (A_n) - (n-1)N \qquad \ldots(11\cdot 4)$$

where N is the total number of observations.

Solution. We have

$$(\alpha_1\alpha_2) = \alpha_1\alpha_2 . N = (1 - A_1)(1 - A_2). N = N - (A_1) - (A_2) + (A_1A_2)$$

Since class frequency is always non-negative, we have

$$(\alpha_1\alpha_2) \geq 0 \Rightarrow (A_1A_2) \geq (A_1) + (A_2) - N \qquad \ldots(*)$$

It follows that (11·4) is true for 2 attributes.

Let us now suppose that (11·4) is true for r attributes $A_1, A_2, \ldots, A_r$ so that

$$(A_1 A_2 A_3 \ldots A_r) \geq (A_1) + (A_2) + (A_3) + \ldots + (A_r) - (r-1)N$$

Replacing the attribute A_r by another compound attribute A_rA_{r+1}, we get

$$\begin{aligned}(A_1 A_2 A_3 \ldots A_rA_{r+1}) &\geq (A_1) + (A_2) + (A_3) + \ldots + (A_rA_{r+1}) - (r-1)N \\ &\geq (A_1) + (A_2) + (A_3) + \ldots + \{(A_r) + (A_{r+1}) - N\} - (r-1)N \\ &\hspace{20em} \text{[From (*)]} \\ &= (A_1) + (A_2) + \ldots + (A_r) + (A_{r+1}) - rN\end{aligned}$$

This implies that if (11·4) is true for $n = r$, it is also true for $n = r + 1$ attributes. But we have seen in (*) that (11·4) is true for $n = 2$. Hence by mathematical induction, the result is true for all positive integral values of n.

Example 11·4. *Show that if A occurs in a larger proportion of the cases where B is than where B is not, then B will occur in a larger proportion of cases where A is than where A is not.*

Solution. The problems can be restated as follows :

Given $\dfrac{(AB)}{(B)} > \dfrac{(A\beta)}{(\beta)}$, prove that $\dfrac{(AB)}{(A)} > \dfrac{(\alpha B)}{(\alpha)}$

Now $\dfrac{(AB)}{(B)} > \dfrac{(A\beta)}{(\beta)} \Rightarrow \dfrac{(\beta)}{(B)} > \dfrac{(A\beta)}{(AB)}$

$\Rightarrow \quad 1 + \dfrac{(\beta)}{(B)} > 1 + \dfrac{(A\beta)}{(AB)}$

$$\Rightarrow \qquad \frac{N}{(B)} > \frac{(A)}{(AB)}$$

$$\Rightarrow \qquad \frac{N}{(A)} > \frac{(B)}{(AB)}$$

$$\Rightarrow \qquad \frac{(A)+(\alpha)}{(A)} > \frac{(AB)+(\alpha B)}{(AB)}$$

$$\Rightarrow \qquad 1+\frac{(\alpha)}{(A)} > 1+\frac{(\alpha B)}{(AB)}$$

$$\Rightarrow \qquad \frac{(\alpha)}{(A)} > \frac{(\alpha B)}{(AB)}$$

$$\Rightarrow \qquad \frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}, \text{ as required.}$$

EXERCISE 11 (a)

1. (*a*) Explain the following : (*i*) Order of a class, (*ii*) Ultimate classes and, (*iii*) Fundamental set of class frequencies.

(*b*) What is meant by a class-frequency of (*i*) first order, (*ii*) third order ? How would you express a class frequency of first order in terms of class frequencies of third order ?

2. What is dichotomy ? Show that the continued dichotomy according to n attributes gives rise to 3^n classes.

3. (*a*) Given that $(AB) = 150$, $(A\beta) = 230$, $(\alpha B) = 260$, $(\alpha\beta) = 2{,}340$; find the other frequencies and the value of N.

(*b*) Given the following frequencies of the positive classes, find the frequencies of the rest of the classes :

$(A) = 977$, $(AB) = 453$, $(ABC) = 127$, $(B) = 1{,}185$, $(AC) = 284$, $N = 12{,}000$, $(C) = 596$, and $(BC) = 250$.

Ans. $(A\beta) = 524$, $(\alpha B) = 732$, $(\alpha\beta) = 10{,}291$, $(\beta\gamma) = 935$, $(\beta C) = 346$, $(\beta\gamma) = 10{,}469$, $(A\gamma) = 693$, $(\alpha C) = 312$, $(AB\gamma) = 326$, $(\alpha BC) = 123$, $(\alpha B\gamma) = 609$, $(A\beta C) = 157$, $(A\beta\gamma) = 367$, $(\alpha\beta C) = 189$, $(\alpha\beta\gamma) = 10{,}192$.

4. Given the following data, find frequencies of (*i*) the remaining positive classes; and (*ii*) the ultimate classes ;

$N = 1{,}800$, $(A) = 850$, $(B) = 780$, $(C) = 326$, $(AB\gamma) = 200$, $(A\beta C) = 94$, $(\alpha BC) = 72$, and $(ABC) = 50$.

5. (*a*) Measurements are made on a thousand husbands and a thousand wives. If the measurements of the husbands exceed the measurements of the wives in 800 cases for one measurement, in 700 cases for another and in 660 cases for both measurements, in how many cases will both measurements on the wife exceed the measurements on the husband ?

Ans. 160

(*b*) An unofficial political study was made about the recent changes in Indian political scene and it was found that 919 Indira Gandhi Congress supporters and 1,268 Organisation Congress supporters wanted socialistic economy, whereas 310 Indira Gandhi Congress supporters and 503 supporters of the Organisation Congress wanted capitalistic economy in the country. Find out the total number of Indira Gandhi's and that of the Organisation's supporters, giving the number of capitalistic economy's and of the socialistic economy's votaries, out of the individuals, who were surveyed.

6. At a competitive examination at which 600 graduates appeared, boys outnumbered girls by 96. Those qualifying for interview exceeded in number those failing to qualify by 310. The number of Science graduate boys interviewed was 300 while among the Arts graduate girls there were 25 who failed to qualify for intervew. Altogether there were only 135 Arts graduates and 33 among them failed to qualify. Boys who failed to qualify numbered 18.

Find (*i*) the number of boys who qualified for interview,

(*ii*) the total number of Science graduate boys appearing, and

(*iii*) the number of Science graduate girls who qualified.

Ans. (*i*) 330, (*ii*) 310, and (*iii*) 53.

7. 100 children took three examinations A, B and C; 40 passed the first, 39 passed the second and 48 passed the third, 10 passed all the three, 21 failed all three, 9 passed the first two and failed the third, 19 failed the first two and passed the third. Find how many children passed at least two examinations. Show that for the question asked certain of the given frequencies are not necessary. Which are they ?

Ans. 38. Only frequencies required are (C), $(\alpha\beta C)$, $(AB\gamma)$.

8. In a university examination, which was indeed very tough, 50% at least failed in "Statistics", 75% at least in Topology, 82% at least in "Functional Analysis" and 96% at least in "Applied Mathematics". How many at least failed in all the four ? (**Ans.** 3%)

Hint. Use the result in Example 11·3. Page 11·6.

9. If a collection contains N items, each of which is characterized by one or more of the attributes A, B, C and D, show that with the usual notations

(*i*) $(ABCD) \geq (A) + (B) + (C) + (D) - 3N$, and

(*ii*) $(ABCD) = (ABD) + (ACD) - (AD) + (AD\beta\gamma)$,

where β and γ represent the characteristics of the absence of B and C respectively.

10. Given $(A) = (\alpha) = (B) = (\beta) = \frac{1}{2}N$; show that $(AB) = (\alpha\beta)$, $(A\beta) = (\alpha B)$.

11. Given that $(A) = (\alpha) = (B) = (\beta) = (C) = (\gamma) = \frac{1}{2}N$

and also that $(ABC) = (\alpha\beta\gamma)$, show that $2(ABC) = (AB) + (AC) + (BC) - \frac{1}{2}N$.

11·6. Consistency of Data. Any class frequencies which have been or might have been observed within one and the same population are said to be consistent if they conform with one another and do not in any way conflict. For example; the figures $(A) = 20$, $(AB) = 25$ are inconsistent as (AB) cannot be greater than (A), if they are observed from the same population.

'*Consistency*' of a set of class frequencies may be defined as the property that none of them is negative, otherwise, the data for class frequencies are said to be 'inconsistent'.

Since any class frequency can be expressed as the sum of some of the ultimate class frequencies, it is necessarily non-negative if all the ultimate class frequencies are non-negative. This provides a criterion for testing the consistency of the data. In fact, we have the following theorem.

Theorem 11·1. "*The necessary and sufficient condition for the consistency of a set of independent class frequencies is that no ultimate class frequency is negative.*"

Remark. We can test the consistency of a set of 2^n algebraically independent class frequencies by calculating the ultimate class frequencies. If any one of them is negative, the given data are inconsistent.

11·6·1. Conditions for consistency of Data. Criteria for consistency of class frequencies are obtained by using theorem 11·1. For a single attribute A we have conditions of consistency as follows :

$$\left.\begin{array}{ll} (i) & (A) \geq 0 \\ (ii) & (\alpha) \geq 0 \Rightarrow (A) \leq N) \end{array}\right\} \quad \ldots(11\cdot5)$$

For two attributes A and B, the conditions of consistency are :

$$\left.\begin{array}{ll} (i) & (AB) \geq 0 \\ (ii) & (A\beta) \geq 0 \Rightarrow (AB) \leq (A) \\ (iii) & (\alpha B) \geq 0 \Rightarrow (AB) \leq (B) \\ (iv) & (\alpha\beta) \geq 0 \Rightarrow (AB) \geq (A) + (B) - N) \end{array}\right\} \quad \ldots(11\cdot6)$$

Conditions of consistency for three attributes A, B and C are

$$\begin{array}{ll} (i) & (ABC) \geq 0 \\ (ii) & (AB\gamma) \geq 0 \Rightarrow (ABC) \leq (AB) \\ (iii) & (A\beta C) \geq 0 \Rightarrow (ABC) \leq (AC) \\ (iv) & (\alpha BC) \geq 0 \Rightarrow (ABC) \leq (BC) \\ (v) & (A\beta\gamma) \geq 0 \Rightarrow (ABC) \geq (AB) + (AC) - (A) \\ (vi) & (\alpha B\gamma) \geq 0 \Rightarrow (ABC) \geq (AB) + (BC) - (B) \\ (vii) & (\alpha\beta C) \geq 0 \Rightarrow (ABC) \geq (AC) + (BC) - (C) \\ (viii) & (\alpha\beta\gamma) \geq 0 \Rightarrow (ABC) \leq (AB) + (BC) + (AC) - (A) - (B) - (C) + N \end{array} \quad \ldots(11\cdot7)$$

(*i*) and (*viii*) in (11·7) give :

$$\left.\begin{array}{lll} & & (AB) + (BC) + (AC) \geq (A) + (B) + (C) - (N) \\ \text{Similarly} & & \\ (ii) \text{ and } (vii) & \Rightarrow & (AC) + (BC) - (AB) \leq (C) \\ (iii) \text{ and } (vi) & \Rightarrow & (AB) + (BC) - (AC) \leq (B) \\ (iv) \text{ and } (v) & \Rightarrow & (AB) + (AC) - (BC) \leq (A) \end{array}\right\} \quad \ldots(11\cdot8)$$

Remark. As already pointed out [*c.f.* Remarks (3) and (4), § 11·4·2)], 2^n algebraically independent class frequencies are necessary to specify the data completely, one such set being the set of ultimate class frequencies and the other being the set of positive class frequencies. If the data supplied are incomplete so that it is not possible to determine all the class frequencies, then the conditions (11·5), (11·6) and (11·8) for one, two and three attributes respectively, enable us to assign the limits within which an unknown class frequency can lie.

Example 11·5. *Examine the consistency of the following data :*

$N = 1{,}000$, $(A) = 600$, $(B) = 500$, $(AB) = 50$, *the symbols having their usual meaning.*

Solution. We have

$$(\alpha\beta) = N - (A) - (B) + (AB) = 1000 - 600 - 500 + 50 = -50.$$

Since $(\alpha\beta) < 0$, the data are inconsistent.

Example 11·6. *Among the adult population of a certain town 50 per cent are males, 60 per cent are wage earners and 50 per cent are 45 years of age or over, 10 per cent of the males are not wage-earners and 40 per cent of the males are under 45. Make the best possible inference about the limits within which the percentage of persons (male or female) of 45 years or over are wage-earners.*

Solution. Let $N = 100$. Then denoting males by A, wage-earners by B and 45 years of age or over by C, we are given :

$$N = 100,\ (A) = 50,\ (B) = 60,\ (C) = 50$$

$$(A\beta) = \frac{10}{100} \times 50 = 5,\ (A\gamma) = \frac{40}{100} \times 50 = 20$$

$$\therefore \quad (AB) = (A) - (A\beta) = 45,\ (AC) = (A) - (A\gamma) = 30$$

We are required to find the limits for (BC).

Conditions of consistency (11·8) give

$$(i) \qquad (AB) + (BC) + (AC) \geq (A) + (B) + (C) - N$$

$$\Rightarrow \qquad (BC) \geq 50 + 60 + 50 - 100 - 45 - 30 = -15$$

$$(ii)\ (AB) + (AC) - (BC) \leq A$$

$$\Rightarrow \qquad (BC) \geq (AB) + (AC) - (A) = 45 + 30 - 50 = 25$$

$$(iii)\ (AB) + (BC) - (AC) \leq (B)$$

$$\Rightarrow \qquad (BC) \leq (B) + (AC) - (AB) = 60 + 30 - 45 = 45$$

$$(iv)\ (AC) + (BC) - (AB) \leq (C)$$

$$(BC) \leq (C) + (AB) - (AC) = 50 + 45 - 30 = 65$$

$$(i) \text{ to } (iv) \Rightarrow \qquad 25 \leq (BC) \leq 45$$

Hence the percentage of wage-earning population of 45 years or over must lie between 25 and 45.

Example 11·7. *In a series of houses actually invaded by smallpox, 70% of the inhabitants are attacked and 85% have been vaccinated. What is the lowest percentage of the vaccinated that must have been attacked ?*

Solution. Let A and B denote the attributes of the inhabitants being attacked and vaccinated respectively. Then we are given :

$$N = 100,\ (A) = 70 \text{ and } (B) = 85$$

Consistency condition gives :

$$(AB) \geq (A) + (B) - N \Rightarrow (AB) \geq 55$$

Hence the lowest percentage of inhabitants vaccinated, who have been attacked is

$$\frac{(AB)}{(B)} \times 100 = \frac{55}{85} \times 100 = 64{\cdot}7\%$$

Example 11·8. *Show that if*

$$\frac{(A)}{N} = x,\ \frac{(B)}{N} = 2x,\ \frac{(C)}{N} = 3x$$

and

$$\frac{(AB)}{N} = \frac{(BC)}{N} = \frac{(CA)}{N} = y,$$

then the value of neither x nor y can exceed 1/4.

Solution. Conditions of consistency give :

$$(AB) \le (A) \Rightarrow Ny \le Nx \Rightarrow y \le x \qquad \ldots(i)$$

Also $(BC) \ge (B) + (C) - N$

$$\Rightarrow \frac{(BC)}{N} \ge \frac{(B)}{N} + \frac{(C)}{N} - 1$$

$$\Rightarrow y \ge 2x + 3x - 1$$

$$\Rightarrow 5x - 1 \le y \qquad \ldots(ii)$$

(*i*) and (*ii*) give

$$5x - 1 \le x \Rightarrow 4x \le 1 \Rightarrow x \le \frac{1}{4} \qquad \ldots(iii)$$

Thus from (*i*) and (*iii*) we have $y \le x \le \frac{1}{4}$, which establishes the result.

Example 11·9. *Show that (i) If all A's are B's and all B's are C's then all A's are C's, (ii) If all A's are Bs and no B's are C's then no A's are C's.*

Solution. (*i*) All A's are B's $\Rightarrow$ $(AB) = (A)$
and all B's are C's $\Rightarrow$ $(BC) = (B)$ $\qquad \ldots(*)$

To prove $(AC) = (A)$

We have $(AB) + (BC) - (AC) \le (B)$

$$\Rightarrow (A) + (B) - (AC) \le (B) \qquad \text{[Using (*)]}$$

$$\Rightarrow (A) \le (AC) \Rightarrow (AC) \ge (A)$$

But since $(AC) \not> (A)$, we have $(AC) = (A)$, as desired.

(*ii*) We are given $(AB) = (A)$ and $(BC) = 0$ and we want to prove $(AC) = 0$.

We have

$$(AB) + (AC) - (BC) \le (A)$$

$$\Rightarrow (A) + (AC) - 0 \le (A)$$

$$\Rightarrow (AC) \le 0$$

And since $(AC) \ge 0$, we must have $(AC) = 0$.

EXERCISE 11(b)

1. What do you understand by consistency of given data ? How do you check it ?

2. (*a*) If a report gives the following frequencies as actually observed, show that there must be a misprint or mistake of some sort, and that possibly the misprint consists in the dropping of 1 before 85 given as the frequency (BC) :

$N = 1000$, $(A) = 510$, $(B) = 490$, $(C) = 427$, $(AB) = 189$, $(AC) = 140$, $(BC) = 85$.

(*b*) A student reported the results of a survey in the following manner, in terms of the usual notations :

$N = 1000$, $(A) = 525$, $(B) = 312$, $(C) = 470$, $(AB) = 42$, $(BC) = 86$, $(AC) = 147$, and $(ABC) = 25$.

Examine the consistency of the above data.

(*c*) Examine the consistency and adequacy of the following data to determine the frequencies of the remaining positive and ultimate classes.

$$N = 10{,}000,\ (A) = 1087,\ (B) = 286,\ (C) = 877,$$
$$(CA\beta) = 281,\ (C\alpha\beta) = 86,\ (\gamma AB) = 78,\ (ABC) = 57$$

3. Given that $(A) = (B) = (C) = \frac{1}{2}N$ and 80 per cent of A's are B's, 75 per cent of A's are C's, find the limits to the percentage of B's that are C's.

Ans. 55% and 95%.

4. If $(A) = 50$, $(B) = 60$, $(C) = 50$, $(A\beta) = 5$, $(A\gamma) = 20$, $N = 100$, find the greatest and the least possible values of (BC) so that the data may be consistent.

Ans. $25 \le (BC) \le 45$

5. If $1{,}000 = N = 1\frac{5}{3}(A) = 2(B) = 2\frac{5}{2}(C) = 5\,(AB)$, and $(AC) = (BC)$, what should be the minimum value of (BC) ?

Ans. 150

6. Given that $(A) = (B) = (C) = \frac{1}{2}N = 50$ and $(AB) = 30$, $(AC) = 25$, find the limits within which (BC) will lie.

7. In a university examination 65% of the candidate passed in English, 90% passed in the second language and 60% passed in the optional subjects. Find how many at least should have passed the whole examination.

Ans. 15%. **Hint.** Use Example 11·3.

8. A market investigator returns the following data. Of 1,000 people consulted 811 liked chocolates, 752 liked toffees and 418 liked boiled sweets, 570 liked both chocolates and toffees, 356 liked chocolates and boiled sweets and 348 liked toffees and boiled sweets, 297 liked all three. Show that this information as it stands must be incorrect.

9. (*a*) In a school, 50 per cent of the students are boys, 60 per cent are Hindus and 50 per cent are 10 years of age or over. Twenty per cent of the boys are not Hindus and 40 per cent of the boys are under 10. What conclusions can you draw in regard to percentage of Hindu students of 10 years or over ?

(*b*) In a college, 50 per cent of the students are boys, 60 per cent of the student are above 18 years and 80 per cent receive scholarships. 35 per cent of the students are boys above 18 years of age, 45 per cent are boys receiving scholarships, and 42 per cent are above 18 years and receive scholarships. Determine the limits to the proportion of boys above 18 years who are in receipt of scholarships.

Ans. Between 30 and 32.

10. The following summary appears in a report on a survey covering 1,000 fields. Scrutinise the numbers and point out if there is any mistake or misprint in them.

Manured fields	510
Irrigated fields	490
Fields growing improved varieties	427
Fields both irrigated and manured	189
Fields both manured and growing improved varieties	140
Fields both irrigated and growing improved varieties	85

Hint. Let A : manured fields;

B : Irrigated fields

and C : Growing improved varieties; then $(\alpha\beta\gamma) < 0$.

11. A social survey in a village revealed that there were more uneducated employed males than educated ones; there were more educated employed males than uneducated unemployed males. There were more educated unemployed under 35 years of age than employed uneducated males over 35 years of age. Show that there are more uneducated employed males under 35 years of age than educated unemployed males over 35 years of age.

12. In a war between White and Red forces, there are more Red soldiers than White, there are more armed Whites than unarmed Reds, there are fewer armed Reds with ammunition than unarmed Whites without ammunition. Show that there are more armed Reds without ammunition than unarmed Whites with ammunition.

13. Given that $(A) = (B) = (C) = \dfrac{N}{2}, \dfrac{(AB)}{N} = \dfrac{(AC)}{N} = p$, find what must be the greatest and least values of p in order that we may infer that $(BC)/N$, exceeds any given value, say q.

Ans. $\frac{1}{4}(1 - 2q) \le p \le \frac{1}{4}(1 + 2q)$.

11·7. Independence of Attributes. Two attributes A and B are said to be independent if there exists no relationship of any kind between them. If A and B are independent, we would expect (*i*) the same proportion of A's amongst B's as amongst β's, (*ii*) the proportion of B's amongst A's is same as that amongst the α's. For example, if insanity and deafness are independent, the proportion of the insane people among deafs and non-deafs must be same.

11·7·1. Criterion of Independence. If A and B are independent, then (*i*) in § 11·7 gives

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \qquad \ldots(11{\cdot}9)$$

$$\Rightarrow \quad 1 - \frac{(AB)}{(B)} = 1 - \frac{(A\beta)}{(\beta)}$$

$$\Rightarrow \quad \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)} \qquad \ldots(11{\cdot}9a)$$

Similarly, (*ii*) in § 11·7 gives

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)} \qquad \ldots(11{\cdot}10)$$

$$\Rightarrow \quad 1 - \frac{(AB)}{(A)} = 1 - \frac{(\alpha B)}{(\alpha)}$$

$$\Rightarrow \qquad \frac{(A\beta)}{(A)} = \frac{(\alpha\beta)}{(\alpha)} \qquad \ldots(11 \cdot 10a)$$

In fact (11·9) ⇒ (11·10) and *vice-versa.*

For example, (11·9) gives

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{(N)}$$

$$\Rightarrow \qquad \frac{(AB)}{(A)} = \frac{(B)}{N} = \frac{(B)-(AB)}{N-(A)} = \frac{(\alpha B)}{(\alpha)}, \qquad (11 \cdot 10b)$$

which is (11·10). Similarly, starting from (11·10), we would arrive at (11·9).

It becomes easier to grasp the nature of the above relations if the frequencies are supposed to be grouped into a table with two rows and two columns as follows:

Attributes	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	N

Second criterion of independence may be obtained in terms of the class frequencies of first order. (11·10*b*) gives

$$(AB) = \frac{(A)\,(B)}{N} \qquad \ldots(11 \cdot 11)$$

$$\Rightarrow \qquad \frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N} \qquad \ldots(11 \cdot 11a)$$

which leads to the following important fundamental rule :

"If the attributes A and B are independent, the proportion of AB's in the population is equal to the product of the proportions of A's and B's in the population."

We may obtain a third criterion of independence in terms of second order class frequencies, as follows.

$$(AB)\,.\,(\alpha\beta) = \frac{(A)\,(B)}{N} \cdot \frac{(\alpha)\,(\beta)}{N} = \frac{(A)\,(\beta)}{N} \cdot \frac{(\alpha)\,(B)}{N} \qquad \text{(Using 11·11)}$$

$$\Rightarrow \qquad \left.\begin{aligned} (AB)\,.\,(\alpha\beta) &= (A\beta)\,.\,(\alpha B) \\ \frac{(AB)}{(\alpha B)} &= \frac{(A\beta)}{(\alpha\beta)} \end{aligned}\right\} \qquad \ldots(11 \cdot 12)$$

Aliter. (11·12) may also be obtained from(11·9) and (11·9*a*) as explained below:

$$(11 \cdot 9) \text{ and } (11 \cdot 9a) \quad \Rightarrow \quad \frac{(AB)}{(A\beta)} = \frac{(B)}{(\beta)} = \frac{(\alpha B)}{(\alpha\beta)}$$

$$\Rightarrow \qquad (AB)(\alpha\beta) = (A\beta) \cdot (\alpha\beta)$$

Similarly, (11·10) and (11·10*a*) give the same result.

11·7·2. Symbols $(AB)_0$ and δ. Let us write

$$(AB)_0 = \frac{(A)\,(B)}{N} \qquad \ldots(11\cdot13)$$

which is the value of (AB) under the hypothesis that the attributes A and B are independent.

Let $$\delta = (AB) - (AB)_0 \qquad \ldots(11\cdot14)$$

denote the excess of (AB) over $(AB)_0$. Then

$$\delta = (AB) - \frac{(A)\,(B)}{N} = \frac{1}{N}\,[N\,(AB) - (A)(B)]$$

$$= \frac{1}{N}\,\Big[\{(AB) + (A\beta) + (\alpha B) + (\alpha\beta)\}\,(AB))$$
$$- \{(AB) + (A\beta)\}\,\{(AB) + (\alpha B)\}\Big]$$

$$= \frac{1}{N}\left[(AB)\,(\alpha\beta) - (A\beta)\,(\alpha B)\right] \qquad \text{[On simplification]}$$

(11·12) $\Rightarrow$ $\delta = 0$, if A and B are independent. ...(11·15)

Example 11·10. *If $\delta = (AB) - (AB)_0$, then with usual notations, prove that*

(*i*) $[(A) - (\alpha)][(B) - (\beta)] + 2N\delta = (AB)^2 + (\alpha\beta)^2 - (A\beta)^2 - (\alpha B)^2$

(*ii*) $\delta = \dfrac{(B)\,(\beta)}{N}\left\{\dfrac{(AB)}{(B)} - \dfrac{(A\beta)}{(\beta)}\right\} = \dfrac{(A)(\alpha)}{N}\left\{\dfrac{(AB)}{(A)} - \dfrac{(\alpha B)}{(\alpha)}\right\}$

Solution. (*i*) We have $\delta = (AB) - (AB)_0 = (AB) - \dfrac{(A)\,(B)}{N}$

$$\text{L.H.S.} = [(A) - (\alpha)][(B) - (\beta)] + 2N\delta$$
$$= [(AB) + (A\beta) - (\alpha B) - (\alpha\beta)][(AB) + (\alpha B) - (A\beta) - (\alpha\beta)]$$
$$+ 2N\left[(AB) - \frac{(A)\,(B)}{N}\right]$$
$$= [\{(AB) - (\alpha\beta)\} + \{(A\beta) - (\alpha B)\}][\{(AB) - (\alpha\beta)\} - \{(A\beta) - (\alpha B)\}]$$
$$+ 2[N(AB) - (A)\,(B)]$$
$$= [(AB) - (\alpha\beta)]^2 - [(A\beta) - (\alpha B)]^2$$
$$+ 2[(AB)\{(AB) + (A\beta) + (\alpha B) + (\alpha\beta)\} - \{(AB) + (A\beta)\}\,\{(AB) + (\alpha B)\}]$$
$$= [(AB)^2 + (\alpha\beta)^2 - 2(AB)\,(\alpha\beta)]$$
$$-[(A\beta)^2 + (\alpha B)^2 - 2(A\beta)\,(\alpha B)] + 2[(AB)\,(\alpha\beta) - (A\beta)\,(\alpha B)]$$

(On simplification)

$$= (AB)^2 + (\alpha\beta)^2 - (A\beta)^2 - (\alpha B)^2 = \text{R.H.S.}$$

(*ii*) $\dfrac{(B)\,(\beta)}{N}\left[\dfrac{(AB)}{(B)} - \dfrac{(A\beta)}{(\beta)}\right] = \dfrac{1}{N}\left[(\beta)(AB) - (B)(A\beta)\right]$

$$= \frac{1}{N}\Big[(AB)\ \{N - (B)\} - (B)\ \{(A) - (AB)\} \Big]$$

$$= \frac{1}{N}\Big[N \,.\, (AB) - (A)\ (B) \Big] = (AB) - \frac{(A)\,.\,(B)}{N} = \delta$$

Since δ is symmetric in A and B, by interchanging A and B, we will obtain the second result.

11·8. Association of Attributes. Two attributes A and B are said to be associated if they are not independent but are related in some way or the other. They are said to be

$$\left.\begin{array}{ll} \textit{positively associated} \text{ if } (AB) > \dfrac{(A)\,(B)}{N} \\ \text{and} \quad \textit{negativel associated} \text{ if } (AB) < \dfrac{(A)\,(B)}{N} \end{array}\right\} \quad \ldots(11\cdot16)$$

In other words, two attributes A and B are positively associated if $\delta > 0$, negatively associated if $\delta < 0$ or and are independent if $\delta = 0$ (c.f. § 11·7·2).

Remarks 1. Two attributes A and B are said to be *completely associated* if A cannot occur without B, though B may occur without A and *vice-versa*. In other words, for complete association either all A's are B's *i.e.*, $(AB) = (A)$ or all B's are A's *i.e.*, $(AB) = (B)$ according as either A's or B's are in a minority. Similarly, *complete dissociation* means that no A's are B's *i.e.*, $(AB) = 0$ or no α's are β's *i.e.*, $(\alpha\beta) = 0$ or more generally when either of these statements is true.

2. It should be carefully noted that the word 'association' used in Statistics is technically different from the general notion of association as used in day-to-day life. Ordinarily, two attributes are said to be associated if they occur together in a number of cases. But statistically two attributes are said to be associated if they occur together in a large number of cases than expected if they were independent, *i.e.*, if $\delta = (AB) - (A)\cdot(B)/N > 0$. In Statistics, the statement that "some A's are B's", however great the proportion, does not necessarily imply association between them. Thus to find out if two attributes are associated, we must know (A), (B), (AB) and N. Incomplete information will not enable us to conclude anything about association between them. For example, consider the following statement :

"90 per cent of the people who drink alcohol die before reaching the age of 75 years. Hence drinking is bad for longevity of life."

The inference drawn is not correct, since the given information is not complete for drawing any valid conclusions about association. It might happen that 95% of the people who do not drink, die before reaching 75 years of age. In that case drinking might be found good for longevity of life.

3. *Sampling fluctuations.* If $\delta \neq 0$ and its value is fairly small, then it is possible that this association is just by chance (or commonly termed as 'due to fluctuations of sampling) and not really significant of any real association between the attributes. We should not, therefore, draw hasty conclusions about association or dissociation unless δ, the difference between (AB) and its expected value (under the hypothesis of independence) $(A)\cdot(B)/N$, is significant. The

problem : 'how much difference is to be regarded as significant' will be discussed in detail in Chapters 12 (Large sample test for attributes) and 13 (Chi-square test of goodness of fit). This point has been raised here only as a precautionary measure to warn the reader against drawing hasty inferences.

11·8·1. Yule's Coefficient of Association. As a measure of the intensity of association between two attributes A and B, G. Udny Yule gave the coefficient of association Q, defined as follows :

$$Q = \frac{(AB)\,(\alpha\beta) - (A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta) + (A\beta)\,(\alpha B)} = \frac{N\delta}{(aB)\,(\alpha\beta) + (A\beta)\,(\alpha B)} \qquad \ldots(11\cdot17)$$

If A and B are *independent*, $\delta = 0 \Rightarrow Q = 0$.

If A and B are *completely associated*, then

either $\quad (AB) = (A) \Rightarrow (A\beta) = 0$

or $\quad (AB) = (B) \Rightarrow (\alpha B) = 0$

and in each case $Q = +1$.

If A and B are in complete dissociation then either $(AB) = 0$ or $(\alpha\beta) = 0$ and we get $Q = -1$.

Hence $\qquad -1 \le Q \le 1 \qquad \ldots(11\cdot18)$

Remark. An important property of Q is that it is independent of the relative proportion of A's or α's in the data. Thus if all the terms containing A in Q are multiplied by a constant, k (say), its value remains unaltered. Similarly for B, β and α. This property renders it specially useful to situations where the proportions are arbitrary, *e.g., experiments.*

11·8·2. Coefficient of Colligation. Another coefficient with the same properties as Q, is the coefficient of colligation Y, given by

$$Y = \left\{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}\right\} \Big/ \left\{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}\right\} \qquad \ldots(11\cdot19)$$

Remarks 1. Obviously $Q = 0 \Rightarrow Y = \dfrac{1-1}{1+1} = 0.$

$Q = -1 \Rightarrow Y = -1$ and $Q = 1 \Rightarrow Y = 1$ and conversely.

2. If we let $\dfrac{(A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta)} = k$, so that

$$Y = \frac{1-\sqrt{k}}{1+\sqrt{k}} \Rightarrow Y^2 = \frac{1+k-2\sqrt{k}}{1+k+2\sqrt{k}}$$

$$\Rightarrow \quad 1 + Y^2 = \frac{2(1+k)}{1+k+2\sqrt{k}} = \frac{2(1+k)}{(1+\sqrt{k})^2}$$

$$\therefore \quad \frac{2Y}{1+Y^2} = \frac{2(1-\sqrt{k})(1+\sqrt{k})}{2(1+k)} = \frac{1-k}{1+k}$$

$$= \frac{1 - \dfrac{(A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta)}}{1 + \dfrac{(A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta)}} = \frac{(AB)\,(\alpha\beta) - (A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta) + (A\beta)\,(\alpha B)}$$

$$\Rightarrow \qquad Q = \frac{2Y}{1+Y^2} \qquad \ldots(11\cdot20)$$

Example 11·11. *Find if A and B are independent, positively associated or negatively associated, in each of the following cases :*

(i) $N = 1000, (A) = 470, (B) = 620,$ *and* $(AB) = 320.$

(ii) $(A) = 490, (AB) = 294, (\alpha) = 570,$ *and* $(\alpha B) = 380.$

(iii) $(AB) = 256, (\alpha B) = 768, (A\beta) = 48,$ *and* $(\alpha\beta) = 144.$

Solution. (*i*) $\delta = (AB) - \dfrac{(A)(B)}{N}$

$$= 320 - \frac{470 \times 620}{1000} = 320 - 291\cdot4 = 28\cdot6$$

Since $\delta > 0$, A and B are positively associated.

(*ii*) We have $N = (A) + (\alpha) = 490 + 570 = 1060$

$$(B) = (AB) + (\alpha B) = 294 + 380 = 674$$

$$\therefore \quad \delta = (AB) - \frac{(A)(B)}{N} = 294 - \frac{490 \times 674}{1060} = 294 - 311\cdot6 < 0$$

Hence A and B are negatively associated.

(*iii*) $(A) = (AB) + (A\beta) = 256 + 48 = 304$

$(B) = (AB) + (\alpha B) = 256 + 768 = 1024$

$N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta) = 256 + 48 + 768 + 144 = 1216$

$$\therefore \quad \delta = (AB) - \frac{(A)(B)}{N} = 256 - \frac{304 \times 1024}{1216} = 0$$

Hence A and B are independent.

Aliter. Since all the four frequencies of order 2 are given, using (11·15), we have

$$\delta = \frac{1}{N}\left[(AB)(\alpha\beta) - (A\beta)(\alpha B)\right] = \frac{1}{N}\left[256 \times 144 - 48 \times 768\right]$$

$$= \frac{256}{N}\left[144 - 48 \times 3\right] = 0$$

$\Rightarrow \qquad A$ and B are independent.

Example 11·12. *Investigate the association between darkness of eye-colour in father and son from the following data :*

Fathers with dark eyes and sons with dark eyes :	*50*
Fathers with dark eyes and sons with not dark eyes :	*79*
Fathers with not dark eyes and sons with dark eyes :	*89*
Fathers with not dark eyes and sons with not dark eyes :	*782*

Also tabulate for comparison the frequencies that would have been observed had there been no heredity.

Solution. Let A : Dark eye-colour of father and

B : Dark eye-colour of son.

Then we are given $(AB) = 50, (A\beta) = 79, (\alpha B) = 89, (\alpha\beta) = 782$

$$\therefore \qquad Q = \frac{50 \times 782 - 79 \times 89}{50 \times 782 + 79 \times 89} = \frac{32069}{46131} = +0.69$$

Hence there is a fairly high degree of positive association between the eye colour of fathers and sons.

We have

$$\begin{aligned}
(A) &= (AB) + (A\beta) = 50 + 79 &&= 129\\
(B) &= (AB) + (\alpha B) = 50 + 89 &&= 139\\
(\alpha) &= (\alpha B) + (\alpha\beta) = 89 + 782 &&= 871\\
(\beta) &= (A\beta) + (\alpha\beta) = 79 + 782 &&= 861\\
N &= (A) + (\alpha) = 129 + 871 &&= 1000
\end{aligned}$$

Under the condition of *no heredity, i.e.,* independence of attributes A and B, we have

$$(AB)_0 = \frac{(A)\,(B)}{N} = \frac{129 \times 139}{1000} = 18\,; \quad (A\beta)_0 = \frac{(A)\,(\beta)}{N} = \frac{129 \times 861}{1000} = 111$$

$$(\alpha B)_0 = \frac{(\alpha)\,(B)}{N} = \frac{871 \times 139}{1000} = 121; \quad (\alpha\beta)_0 = \frac{(\alpha)\,(\beta)}{N} = \frac{871 \times 861}{1000} = 750$$

Example 11·13. *Can vaccination be regarded as a preventive measure for small pox from the data give below ?*

'Of 1482 persons in a locality exposed to small-pox, 368 in all were attacked.'

'Of 1482 persons; 343 had been vaccinated and of these only 35 were attacked.'

Solution. Let A denote the attribute of vaccination and B that of attack by small-pox. Then the given data are :

$$N = 1482,\ (A) = 368,\ (B) = 343 \ \text{ and } \ (AB) = 35$$

$$(\alpha\beta) = N - (A) - (B) + (AB) = 1482 - 368 - 343 + 35 = 806$$

$$(A\beta) = (A) - (AB) = 368 - 35 = 333$$

$$(\alpha B) = (B) - (AB) = 343 - 35 = 308$$

$$\therefore \qquad Q = \frac{(AB)\,(\alpha\beta) - (A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta) + (a\beta)\,(\alpha B)} = \frac{35 \times 806 - 333 \times 308}{35 \times 806 + 333 \times 308} = -0.57$$

Thus, there is negative association between A and B *i.e.,* between 'attacked' and 'vaccinated'. In other words, there is positive association between not attacked and vaccinated. Hence vaccination can be regarded as a preventive measure for smallpox.

EXERCISE 11(c)

1. (*a*) What do you mean by independence of attributes ? Give a criterion of independence for attributes A and B.

(*b*) What are the various methods of finding whether two attributes are associated, dissociated or independent ? Deduce any one such measure of association.

(*c*) When are two attributes said to be positively associated and negatively associated ? Also define complete association and dissociation of two attributes.

(*d*) Derive an expression for a measure of association between two attributes.

(*e*) What is association of attributes ? Write a note on the strength of association and how it is measured ?

(*f*) Find whether the attributes α and β are positively associated, negatively associated or independent. Given $(AB) = 500$, $(\alpha) = 800$, $(B) = 600$, $N = 1500$.

2. (*a*) Define Yule's coefficient of association and the coefficient of Colligation. Establish the following relation between coefficient of association Q and coefficient of colligation Y :

$$Q = \frac{2Y}{1 + Y^2}$$

(*b*) For the following table, give Yule's coefficient of association (Q) and coefficient of Colligation (Y). Examine the cases (*i*) $bc = 0$, (*ii*) $ad = 0$, and (*iii*) $ad = bc$.

	B	not B
A	a	b
not A	c	d

Ans. $Q = 1 = Y$ if $bc = 0$ and $Q = -1 = Y$ if $ad = 0$ and $Q = 0$ if $ad = bc$.

(*c*) Prove that in the usual notations $Q = 2Y/(1 + Y^2)$. What is the range of values for Q ?

(*d*) If an attribute A is known to be completely associated with an attribute B, (*i*) what can you infer about the association between α and β ? (α and β are equivalent to 'not A' and 'not B' respectively), (*ii*) α and B ?

3. (*a*) The following table is reproduced from a memoir written by Karl Pearson :

		Eye colour in son	
		Not light	Light
Eye colour in father	Not light	230	148
	Light	151	471

Discuss if the colour of son's eyes is associated with that of father.

Ans. Yes. Positively associated, $Q = 0{\cdot}66$.

(*b*) The following table shows the result of inocculation against cholera.

	Not attacked	Attacked
Inocculated	431	5
Not-inocculated	291	9

Examine the effect of inocculation in controlling susceptibility to cholera.

Ans. Inocculation is effective in controlling cholera.

4. (*a*) find the association between proficiency in English and in Hindi among candidates at a certain test if 245 of them passed in Hindi, 285 failed in Hindi, 190 failed in Hindi but passed in English and 147 passed in both.

(*b*) The male population of a state is 250 lakhs. The number of literate males is 20 lakhs and total number of male criminals is 26 thousand. The number of literate male criminals is 2 thousand. Do you find any association between literacy and criminality ?

Ans. Literacy and criminality are positively associated.

(c) From the following particulars find whether blindness and baldness are associated :

Total population	1,62,64,000
Number of baldheaded	24,441
Number of blind	7,263
Number of baldheaded blind	221

5. In a certain investigation carried on with regard to 500 graduates and 1500 non-graduates, it was found that the number of employed graduates was 450 while the number of unemployed non-graduates was 300. In the second investigation 5000 cases were examined. The number of non-graduates was 3000 and the number of employed non-graduates was 2500. The number of graduates who were found to be employed was 1600.

Calculate the coefficient of association between graduation and employment in both the investigations.

Can any definite conclusion be drawn from the coefficients ?

Ans. Q (1st Investigation) = +0·38, Q (Second Investigation) = – 0·11

6. (a) Three aptitude tests A, B, C were given to 200 apprentice trainees. From amongst them 80 passed test A, 78 passed test B and 96 passed the third test. While 20 passed all the three tests, 42 failed all the three, 18 passed A and B but failed C and 38 failed A and B but passed the third. Determine (i) how many trainees passed at least two of the three tests and (ii) whether the performances in tests A and B are associated. **Ans.** (i) 76, (ii) $Q = 0{\cdot}3$

(b) In a survey of a population of 12000, information is gathered regarding three attributes A, B and C. In the usual notations,

$$(A) = 980;\ (AB) = 450,\ (ABC) = 130$$
$$(B) = 1190,\ (AC) = 280,\ (C) = 600 \text{ and } (BC) = 250.$$

Find : (i) $(\alpha\beta\gamma)$ (ii) Q_{AB} = Coefficient of Association between A and B

Comment on your findings.

7. A group of 1000 fathers was studied and it was found that 12·9% had dark eyes. Among them the ratio of those having sons with dark eyes to those having sons with not dark eyes was 1 : 1·58. The number of cases where fathers and sons both did not have dark eyes was 782. Calculate coefficient of association between darkness of eye colour in father and son. Give the frequencies that would have been observed had there been completely no heredity.

Hint. $(AB) = 50$, $(A\beta) = 79$, $(\alpha B) = 89$ and $(\alpha\beta) = 782$.

8. A census revealed the following figures of the blind and the insane in two age-groups in a certain population :

	Age-Group 15—25 years	*Age-Group over 25 years*
Total population	2,70,000	1,60,200
Number of blind	1,000	2,000
Number of insane	6,000	1,000
Number of insane among the blind	19	9

(i) Obtain a measure of association between blindness and insanity for each age-group.

(*ii*) Which group shows more association or dis-association (if any) ?

9. Show that if $(AB)_1$, $(\alpha B)_1$, $(A\beta)_1$, $(\alpha\beta)_1$ and $(AB)_2$, $(\alpha B)_2$, $(A\beta)_2$ and $(\alpha\beta)_2$ be two aggregates corresponding to the same values of (A), (B), (α) and (β), then

$$(AB)_1 - (AB)_2 = (\alpha B)_2 - (\alpha B)_1 = (A\beta)_2 - (A\beta)_1 = (\alpha\beta)_1 - (\alpha\beta)_2$$

10. Show that if $\delta = (AB) - \dfrac{(A)(B)}{N}$, then

$$\delta = \frac{1}{N}\left[(AB)(\alpha\beta) - (A\beta)(\alpha B)\right]$$

OBJECTIVE TYPE QUESTIONS

I. State, giving reasons, whether each of the following statements is true or false :

(*i*) There is no difference between correlation and association.

(*ii*) All the class frequencies of various orders are independent of each other.

(*iii*) If the attributes A and B are positively associated, then α and B are also positively associated.

(*iv*) Square of Yule's coefficient of association cannot exceed 1.

(*v*) Yule's coefficient of association cannot be negative.

(*vi*) For two attributes A and B, the coefficient of association Q is 0·36. If each ultimate class frequency is doubled then Q is 0·72.

(*vii*) If $(AB) = 10$, $(\alpha B) = 15$, $(A\beta) = 20$ and $(\alpha\beta) = 30$, then A and B are associated.

(*viii*) If every item which possesses an attribute A possesses the attribute B as well, then the coefficient of association between A and B is 1.

II. Indicate the correct answer :

(*i*) In case of two attributes A and B, the ultimate class frequencies are :

(*a*) : (A), (*b*) : (AB), (*c*) : (α), (*d*) : (B).

(*ii*) The condition for the consistency of a set of independent class frequencies is that no ultimate class frequency is (*a*) zero, (*b*) positive, (*c*) negative.

(*iii*) Attributes A and B are said to be independent if

(*a*) $(AB) > \dfrac{(A)\times(B)}{N}$, (*b*) $(AB) = \dfrac{(A)\times(B)}{N}$. (*c*) $(AB) < \dfrac{(A)\times(B)}{N}$

(*iv*) Attributes A and B are said to be positively associated if

(*a*) $\dfrac{(AB)}{(B)} < \dfrac{(A\beta)}{(\beta)}$, (*b*) $\dfrac{(AB)}{(B)} = \dfrac{(A\beta)}{(\beta)}$, (*c*) $\dfrac{(AB)}{(B)} > \dfrac{(A\beta)}{(\beta)}$, (*d*) $\dfrac{(AB)}{(A)} < \dfrac{(A\beta)}{(\beta)}$

(*v*) If $N = 50$, $(A) = 35$, $(B) = 25$, $(AB) = 15$, then the attributes A and B are said to be :

(*a*) correlated, (*b*) independent, (*c*) negatively associated, (*d*) positively associated

(*vi*) When there is a perfect positive association between two attributes, Q would be (*a*) zero, (*b*) $-0{\cdot}9$, (*c*) -1, (*d*) $+1$.

CHAPTER TWELVE

Sampling and Large Sample Tests

12·1. Sampling—Introduction. Before giving the notion of sampling we will first define *population*. In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called *population or universe*. Thus in statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.

It is obvious that for any statistical investigation complete enumeration of the population is rather impracticable. For example, if we want to have an idea of the average per capita (monthly) income of the people in India, we will have to enumerate all the earning individuals in the country, which is rather a very difficult task.

If the population is infinite, complete enumeration is not possible. Also if the units are destroyed in the course of inspection (*e.g.*, inspection of crackers, explosive materials, etc.), 100% inspection, though possible, is not at all desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse to because of multiplicity of causes, *viz.*, administrative and financial implications, time factor, etc., and we take the help of *sampling*.

A finite subset of statistical individuals in a population is called a *sample* and the number of individuals in a sample is called the sample size.

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the sample only are observed. Then the sample characteristics are utilised to approximately determine or estimate the population. For example, on examining the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff. The error involved in such approximation is known as *sampling error* and is inherent and unavoidable in any and every sampling scheme. But sampling results in considerable gains, especially in time and cost not only in respect of making observations of characteristics but also in the subsequent handling of the data.

Sampling is quite often used in our day-to-day practical life. For example, in a shop we assess the quality of sugar, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

12·2. Types of Sampling. Some of the commonly known and frequently used types of sampling are :

(*i*) *Purposive sampling*, (*ii*) *Random sampling*, (*iii*) *Stratified sampling*, (*iv*) *Systematic Sampling*.

Below we will precisely explain these terms, without entering into detailed discussion.

12·2·1. Purposive Sampling. Purposive sampling is one in which the sample units are selected with definite purpose in view. For example, if we want to give the picture that the standard of living has increased in the city of New Delhi, we may take individuals in the sample from rich and posh localities like Defence Colony, South Extension, Golf Links, Jor Bagh, Chanakyapuri, Greater Kailash etc. and ignore the localities where low income group and the middle class families live. This sampling suffers from the drawback of favouritism and nepotism and does not give a representative sample of the population.

12·2·2 Random Sampling. In this case the sample units are selected at random and the drawback of purposive sampling, *viz.*, favouritism or subjective element, is completely overcome. A *random sample* is one in which each unit of population has an equal chance of being included in it.

Suppose we take a sample of size n from a finite population of size N. Then there are ${}^NC_n$ possible samples. A sampling technique in which each of the ${}^NC_n$ samples has an equal chance of being selected is known as *random sampling* and the sample obtained by this technique is termed as a *random sample.*

Proper care has to be taken to ensure that the selected sample is random. Human bias, which varies from individual to individual, is inherent in any sampling scheme administered by human beings. Fairly good random samples can be obtained by the use of *Tippet's random number tables* or by throwing of a dice, draw of a lottery, etc.

The simplest method, which is normally used, is the *lottery system* which is illustrated below by means of an example.

Suppose we want to select 'r' candidates out of n. We assign the numbers one to n, one number to each candidate and write these numbers (1 to n) on n slips which are made as homogeneous as possible in shape, size, etc. These slips are then put in a bag and thoroughly shuffled and then 'r' slips are drawn one by one. The 'r' candidates corresponding to the numbers on the slips drawn, will constitute the random sample.

Remark. *Tippet's Random Numbers.* L.H.C. Tippet's random numbers tables consist of 10400 four-digited numbers, giving in all 10400×4, *i.e.*, 41600 digits, taken from the British census reports. These tables have proved to be fairly random in character. Any page of the table is selected at random and the number in any row or column or diagonal selected at random may be taken to constitute the sample.

12·2·3. Simple Sampling. Simple sampling is random sampling in which each unit of the population has an equal chance, say p, of being included in the sample and that this probability is independent of the previous drawings. Thus a simple sample of size n from a population may be identified with a series of n independent trials with constant probability 'p' of success for each trial.

Remark. It may be pointed out that random sampling does not necessarily imply simple sampling though, obviously, the converse is true. For example, if an urn contains 'a' white balls and 'b' black balls, the probability of drawing a white ball at the first draw is $[a/(a+b)] = p_1$, (say) and if this ball is not replaced the probability of getting a white ball in the second draw is $[(a-1)(a+b-1)] = p_2 \neq p_1$, the sampling is not simple. But since in the first draw each white ball has the same chance, *viz.*, $a/(a+b)$, of being drawn and in the second draw again each white ball has the same chance, *viz.*, $(a-1)/(a+b-1)$, of being drawn, the sampling is random. Hence in this case, the sampling, though random, is not simple. To ensure that sampling is simple, it must be done with replacement, if population is finite. However, in case of infinite population no replacement is necessary.

12·2·4. Stratified Sampling. Here the entire heterogeneous population is divided into a number of homogeneous groups, usually termed as *strata*, which differ from one another but each of these groups is homogeous within itself. Then units are sampled at random from each of these stratum, the sample size in each stratum varies according to the relative importance of the stratum in the population. The sample, which is the aggregate of the sampled units of each of the stratum, is termed as *stratified sample* and the technique of drawing this sample is known as *stratified sampling*. Such a sample is by far the best and can safely be considered as representative of the population from which it has been drawn.

12·3. Parameter and Statistic. In order to avoid verbal confusion with the statistical constants of the population, *viz.*, mean (μ), variance σ^2, etc., which are usually referred to as *parameters*, statistical measures computed from the sample observations alone, *e.g.*, mean ($\bar{x}$), variance (s^2), etc., have been termed by Professor R.A. Fisher as *statistics*.

In practice, parameter values are not known and the estimates based on the sample values are generally used. Thus statistic which may be regarded as an estimate of parameter, obtained from the sample, is a function of the sample values only. It may be pointed out that a statistic, as it is based on sample values and as there are multiple choices of the samples that can be drawn from a population, varies from sample to sample. The determination or the characterisaton of the variation (in the values of the statistic obtained from different samples) that may be attributed to chance or fluctuations of sampling is one of the fundamental problems of the sampling theory.

Remarks 1. Now onwards, μ and σ^2 will refer to the population mean and variance respectively while the sample mean and variance will be denoted by $\bar{x}$ and s^2 respectively.

2. Unbiased Estimate. A statistic $t = t(x_1, x_2, \ldots, x_n)$, a function of the sample values $x_1, x_2, \ldots, x_n$ is an unbiased estimate of population parameter θ, if $E(t) = \theta$. In other words, if

$$E(Statistic) = \text{Parameter}, \qquad \ldots(12\cdot1)$$

then statistic is said to be an unbiased estimate of the parameter.

12·3·1. Sampling Distribution of a Statistic. If we draw a sample of size n from a given finite population of size N, then the total number of possible samples is :

$$^NC_n = \frac{N!}{n!\,(N-n)!} = k, \ (say).$$

For each of these k samples we can compute some statistic $t = t(x_1, x_2, \ldots, x_n)$, in particular the mean $\bar{x}$, the variance s^2, etc., as given below :

Sample Number	*Statistics*		
	t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
k	t_k	$\bar{x}_k$	s_k^2

The set of the values of the statistic so obtained, one for each sample, constitutes what is called the *sampling distribution* of the statistic. For example, the values $t_1, t_2, t_3, \ldots, t_k$ determine the sampling distribution of the statistic t. In other words, statistic t may be regarded as a random variable which can take the values $t_1, t_2, t_3, \ldots, t_k$ and we can compute the various statistical constants like mean, variance, skewness, kurtosis etc., for its distribution. For example, the mean and variance of the sampling distribution of the statistic t are given by :

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + \ldots + t_k) = \frac{1}{k}\sum_{i=1}^{k} t_i$$

$$\text{Var}(t) = \frac{1}{k}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \ldots + (t_k - \bar{t})^2\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2$$

12·3·2. Standard Error. The standard deviation of the sampling distribution of a statistic is known as its *Standard Error*, abbreviated as S.E. The standard errors of some of the well known statistics, *for large samples*, are given below, where n is the sample size, σ^2 the population variance, and P the population proportion, and $Q = 1 - P$, n_1 and n_2 represent the sizes of two independent random samples respectively drawn from the given population(s).

S.No.	*Statistic*	*Standard Error*
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion 'p'	$\sqrt{PQ/n}$
3.	Sample s.d. : s	$\sqrt{\sigma^2/2n}$
4.	Sample variance : s^2	$\sigma^2\sqrt{2/n}$
5.	Sample quartiles	$1{\cdot}36263\ \sigma/\sqrt{n}$
6.	Sample median	$1{\cdot}25331\ \sigma/\sqrt{n}$

7.	Sample correlation coefficient (r)	$(1-\rho^2)/\sqrt{n}$, ρ being the population correlation coefficient
8.	Sample moment μ_3	$\sigma^3\sqrt{96/n}$
9.	Sample moment μ_4	$\sigma^4\sqrt{96/n}$
10.	Sample coefficient of variation (v)	$\dfrac{v}{\sqrt{2n}}\sqrt{1+\dfrac{2v^3}{10^4}} \simeq \dfrac{v}{\sqrt{2n}}$
11.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$
12.	Difference of two sample s.d.'s : $(s_1 - s_2)$	$\sqrt{\dfrac{\sigma_1^2}{2n_1}+\dfrac{\sigma_2^2}{2n_2}}$
13.	Difference of two sample proportions $(p_1 - p_2)$	$\sqrt{\dfrac{P_1Q_1}{n_1}+\dfrac{P_2Q_2}{n_2}}$

Remark on the Utility of Standard Error. S.E. plays a very important role in the large sample theory and forms the basis of the testing of hypothesis. If t is any statistic, then for large samples

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0, 1) \qquad (c.f. \S\ 12{\cdot}9)$$

$$\Rightarrow \qquad Z = \frac{t - E(t)}{\text{S.E. }(t)} \sim N(0, 1), \text{ for large samples.}$$

Thus, if the discrepancy between the observed and the expected (hypothetical) value of a statistic is greater than z_α (c.f. § 12·7·2) times its S.E., the null hypothesis is rejected at α level of significance. Similarly, if

$$| t - E(t) | \leq z_\alpha \times \text{S.E. } (t),$$

the deviation is not regarded significant at 5% level of significance. In other words, the deviation, $t - E(t)$, could have arisen due to fluctuations of sampling and the data do not provide us any evidence against the null hypothesis which may, therefore, be accepted at α level of significance. [For details see § 12·7·3]

(*i*) The magnitude of the standard error gives an index of the precision of the estimate of the parameter. The reciprocal of the standard error is taken as the measure of reliability or precision of the statistic.

$$\text{S.E. }(p) = \sqrt{PQ/n} \qquad [c.f. \ (4b)\ \S\ 12{\cdot}9{\cdot}1]$$

and

$$\text{S.E. }(\bar{x}) = \sigma/\sqrt{n} \qquad [c.f. \ \S\ 12{\cdot}2]$$

In other words, the standard errors of p and $\bar{x}$ vary inversely as the square root of the sample size. Thus in order to double the precision, which amounts to reducing the standard error to one half, the sample size has to be increased four times.

(*ii*) S.E. enables us to determine the probable limits within which the population parameter may be expected to lie. For example, the probable limits for population proportion P are given by

$$p \pm 3\sqrt{pq/n} \qquad (c.f. \text{ Remark § 12·9·1})$$

Remark. S.E. of a statistic may be reduced by increasing the sample size but this results in corresponding increase in cost, labour and time, etc.

12·4. Tests of Significance. A very important aspect of the sampling theory is the study of the *tests of significance*, which enable us to decide on the basis of the sample results, if

(*i*) the deviation between the observed sample statistic and the hypothetical parameter value, or

(*i*) the deviation between two independent sample statistics;

is significant or might be attributed to chance or the fluctuations of sampling.

Since, for large n, almost all the distributions, *e.g.*, Binomial, Poisson, Negative binomial, Hypergeometric (*c.f.* Chapter 7), t, F (Chapter 14), Chi-square (Chapter 13), can be approximated very closely by a normal probability curve, we use the *Normal Test of Significance* (*c.f.* § 12·9) for large samples. Some of the well known tests of significance for studying such differences for small samples are *t-test, F-test* and Fisher's *z-transformation.*

12·5. Null Hypothesis. The technique of randomisation used for the selection of sample units makes the test of significance valid for us. For applying the test of significance we first set up a hypothesis—a definite statement about the population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called *null hypothesis* and is usually denoted by H_0. According to *Prof. R.A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.*

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics, H_0 will be that the sample statistics do not differ significantly.

Having set up the null hypothesis we compute the probability P that the deviation between the observed sample statistic and the hypothetical parameter value might have occurred due to fluctuations of sampling (*c.f.* § 12·7). If the deviation comes out to be significant (as measured by a test of significance), null hypothesis is refuted or rejected at the particular level of significance adopted (*c.f.* § 12·7) and if the deviation is not significant, null hypothesis may be retained at that level.

12·5·1. Alternative Hypothesis. Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1. For example, if we want to test the null hypothesis that the population has a specified mean μ_0, (say), *i.e.*, $H_0 : \mu = \mu_0$, then the alternative hypothesis could be

(*i*) $H_1 : \mu \neq \mu_0$ (*i.e.*, $\mu > \mu_0$ or $\mu < \mu_0$)

(*ii*) $H_1 : \mu > \mu_0$

(*iii*) $H_1 : \mu < \mu_0$

The alternative hypothesis in (*i*) is known as a *two tailed alternative* and the alternatives in (*ii*) and (*iii*) are known as *right tailed* and *left-tailed alternatives* respectively. The setting of alternative hypothesis is very important since it enables us to decide whether we have to use a single-tailed (right or left) or two-tailed test [c.f. § 12·7·1].

12·6. Errors in Sampling. The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice we decide to accept or reject the lot after examining a sample from it. As such we are liable to commit the following two types of errors :

Type I Error : *Reject H_0 when it is true.*

Type II Error : *Accept H_0 when it is wrong, i.e., accept H_0 when H_1 is true.*

If we write.

$$\left.\begin{aligned} P\{\text{Reject } H_0 \text{ when it is true}\} &= P\{\text{Reject } H_0 \mid H_0\} = \alpha \\ \text{and } P\{\text{Accept } H_0 \text{ when it is wrong}\} &= P\{\text{Accept } H_0 \mid H_1\} = \beta \end{aligned}\right\} \quad ...(12\cdot2)$$

then α and β are called the *sizes of type I error and type II error*, respectively.

In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad.

$$\left.\begin{aligned} \text{Thus} \quad & P\{\text{Reject a lot when it is good}\} = \alpha \\ \text{and} \quad & P\{\text{Accept a lot when it is bad}\} = \beta \end{aligned}\right\} \quad ...(12\cdot2a)$$

where α and β are referred to as *Producer's risk* and *Consumer's risk*, respectively.

12·7. Critical Region and Level of Significance. A region (corresponding to a statistic t) in the sample space S which amounts to rejection of H_0 is termed as *critical region* or *region of rejection.* If ω is the critical region and if $t = t(x_1, x_2, ..., x_n)$ is the value of the statistic based on a random sample of size n, then

$$P(t \in \omega \mid H_0) = \alpha, \; P(t \in \bar{\omega} \mid H_1) = \beta \quad (12\cdot2b)$$

where $\bar{\omega}$, the complementary set of ω, is called the *acceptance region.*

We have $\omega \cup \bar{\omega} = S$ and $\omega \cap \bar{\omega} = \phi$

The probability 'α' that a random value of the statistic t belongs to the critical region is known as the *level of significance.* In other words, level of significance is the size of the type I error (or the maximum producer's risk). The levels of significance usually employed in testing of hypothesis are 5% and 1%. The level of significance is always fixed in advance before collecting the sample information.

12·7·1. One tailed and Two Tailed Tests. In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called a *one tailed test.* For example, a test for testing the mean of a population

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis :

$H_1 : \mu > \mu_0$ (Right tailed) or $H_1 : \mu < \mu_0$ (Left tailed), is a *single tailed test*. In the right tailed test ($H_1 : \mu_1 > \mu_0$), the critical region lies entirely in the right tail of the sampling distribution of $\bar{x}$, while for the left tail test ($H_1 : \mu < \mu_0$), the critical region is entirely in the left tail of the distribution.

A test of statistical hypothesis where the alternative hypothesis is two tailed such as :

$H_0 : \mu = \mu_0$, against the alternative hypothesis $H_1 : \mu \neq \mu_0$, ($\mu > \mu_0$ and $\mu < \mu_0$),

is known as *two tailed test* and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.

In a particular problem, whether one tailed or two tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two-tailed we apply two-tailed test and if alternative hypothesis is one-tailed, we apply one tailed test.

For example, suppose that there are two population brands of bulbs, one manufactured by standard process (with mean life μ_1) and the other manufactured by some new technique (with mean life μ_2). If we want to test if the bulbs differ significantly, then our null hypothesis is $H_0 : \mu_1 = \mu_2$ and alternative will be $H_1 : \mu_1 \neq \mu_2$, thus giving us a two-tailed test. However, if we want to test if the bulbs produced by new process have higher average life than those produced by standard process, then we have

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 < \mu_2,$$

thus giving us a left-tail test. Similarly, for testing if the product of new process is inferior to that of standard process, then we have :

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 > \mu_2,$$

thus giving us a right-tail test. Thus, the decision about applying a two-tail test or a single-tail (right or left) test will depend on the problem under study.

12·7·2. Critical Values or Significant Values. The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the *critical value* or *significant* value. It depends upon :

(*i*) The level of significance used, and

(*ii*) The alternative hypothesis, whether it is two-tailed or single-tailed.

As has been pointed out earlier, for large samples, the standardised variable corresponding to the statistic *t viz.* :

$$Z = \frac{t - E(t)}{S.E.(t)} \sim N(0, 1), \qquad \ldots(*)$$

asymptotically as $n \to \infty$. The value of Z given by (*) under the null hypothesis is known as *test statistic*. The critical value of the test statistic at level of significance α for a two-tailed test is given by z_α where z_α is determined by the equation

$$P(|Z| > z_\alpha) = \alpha \qquad \ldots(12\cdot2c)$$

i.e., z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is a symmetrical curve, from (12·2c), we get

$$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha \qquad \text{[By symmetry]}$$

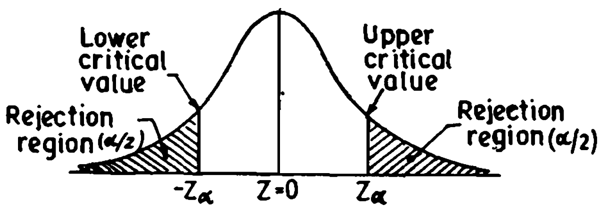
$$\Rightarrow \qquad P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha$$

$$\Rightarrow \qquad 2P(Z > z_\alpha) = \alpha$$

$$\Rightarrow \qquad P(Z > z_\alpha) = \frac{\alpha}{2}$$

i.e., the area of each tail is $\alpha/2$. Thus z_α is the value such that area to the right of z_α is $\alpha/2$ and to the left of $-z_\alpha$ is $\alpha/2$, as shown in the following diagram.

TWO-TAILED TEST
(Level of Significance 'α')

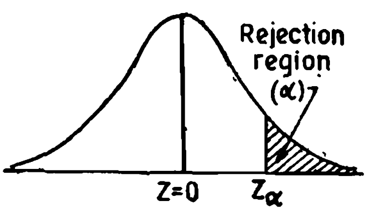


In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $-z_\alpha$ is α (See diagrams below), *i.e.*,

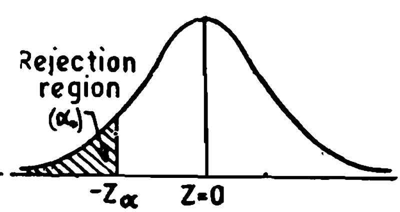
For Right-tail Test : $P(Z > z_\alpha) = \alpha$ ...(12·2d)

For Left-tail Test : $P(Z < -z_\alpha) = \alpha$ ...(12·2e)

RIGHT-TAILED TEST
(Level of Signifiance 'α')



LEFT-TAILED TEST
(Level of Significance 'α')



Thus *the significant* or *critical value of Z for a single-tailed test (left or right) at level of significance 'α' is same as the critical value of Z for a two-tailed test at level of significance '2α'.*

We give on page 12·10, the critical values of Z at commonly used levels of significance for both two-tailed and single-tailed tests. These values have been obtained from equations (12·2c), (12·2d) and (12·2e), on using the Normal Probability Tables as explained in § 12·8.

CRITICAL VALUES (z_α) OF Z

Critical Values (z_α)	Level of significance (α) 1%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2{\cdot}58$	$\lvert Z_\alpha \rvert = 1{\cdot}96$	$\lvert Z_\alpha \rvert = 1{\cdot}645$
Right-tailed test	$Z_\alpha = 2{\cdot}33$	$Z_\alpha = 1{\cdot}645$	$Z_\alpha = 1{\cdot}28$
Left-tailed test	$Z_\alpha = -2{\cdot}33$	$Z_\alpha = -1{\cdot}645$	$Z_\alpha = -1{\cdot}28$

Remark. If n is small, then the sampling distribution of the test statistic Z will not be normal and in that case we can't use the above significant values, which have been obtained from normal probability curves. In this case, *viz.*, n small, (usually less than 30), we use the significant values based on the exact sampling distribution of the statistic Z, [defined in (*), § 12·7·2], which turns out to be t, F, or χ^2 [see Chapters 13, 14]. These significant values have been tabulated for different values of n and α and are given in the Appendix at the end of the book.

12·7·3. Procedure for Testing of Hypothesis. We now summarise below the various steps in testing of a statistical hypothesis in a systematic manner.

1. *Null Hypothesis.* Set up the Null Hypothesis H_0 (see § 12·5, page 12·6).

2. *Alternative Hypothesis.* Set up the Alternative Hypothesis H_1. Ths will enable us to decide whether we have to use a single-tailed (right or left) test or two-tailed test.

3. *Level of Significance.* Choose the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn, *i.e.*, α is fixed in advance.

4. *Test Statistic (or Test Criterion).* Compute the test statistic

$$Z = \frac{t - E(t)}{S.E.(t)}$$

under the null hypothesis.

5. *Conclusion.* We compare z the computed value of Z in step 4 with the significant value (tabulated value) z_α, at the given level of significance, 'α'.

If $\lvert Z \rvert < z_\alpha$, *i.e.*, if the calculated value of Z (in modulus value) is less than z_α we say it is not significant. By this we mean that the difference $t - E(t)$ is just due to fluctuations of sampling and the sample data do not provide us sufficient evidence against the null hypothesis which may therefore, be accepted.

If $\lvert Z \rvert > z_\alpha$, *i.e.*, if the computed value of test statistic is greater than the critical or significant value, then we say that it is significant and the null hypothesis is rejected at level of significance α *i.e.*, with confidence coefficient $(1 - \alpha)$.

12·8. Test of Significance for Large Samples. In this section we will discuss the tests of significance when samples are large. We have seen that for large values of n, the number of trials, almost all the distributions, *e.g.*, binomial, Poisson, negative binomial, etc., are very closely approximated by normal distribution. Thus in this case we apply the *normal test,* which is based upon the following fundamental property (*area property*) of the normal probability curve.

If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma} = \dfrac{X - E(X)}{\sqrt{V(X)}} \sim N(0, 1)$

Thus from the normal probability tables, we have

$$P(-3 \le Z \le 3) = 0.9973, \quad i.e., \quad P(|Z| \le 3) = 0.9973$$

$$\Rightarrow \quad P(|Z| > 3) = 1 - P(|Z| \le 3) = 0.0027 \qquad \ldots(12\cdot3)$$

i.e., in all probability we should expect a standard normal variate to lie between ± 3.

Also from the normal probability tables, we get

$$P(-1.96 \le Z \le 1.96) = 0.95 \quad i.e., \quad P(|Z| \le 1.96) = 0.95$$

$$\Rightarrow \quad P(|Z| > 1.96) = 1 - 0.95 = 0.05 \qquad \ldots(12\cdot3a)$$

and $\quad P(|Z| \le 2.58) = 0.99$

$$\Rightarrow \quad P(|Z| > 2.58) = 0.01 \qquad \ldots(12\cdot3b)$$

Thus the significant values of Z at 5% and 1% level of significance for a two tailed test are 1·96 and 2·58 respectively.

Thus the steps to be used in the normal test are as follows :

(*i*) Compute the test statistic Z under H_0.

(*ii*) If $|Z| > 3$, H_0 is always rejected.

(*iii*) If $|Z| \le 3$, we test its signficance at certain level of significance, usually at 5% and sometimes at 1% level of significance. Thus, for a two-tailed test if $|Z| > 1.96$, H_0 is rejected at 5% level of significance.

Similarly if $|Z| > 2.58$, H_0 is contradicted at 1% level of significance and if $|Z| \le 2.58$, H_0 may be accepted at 1% level of significance.

From the normal probability tables, we have :

$$\begin{aligned} P(Z > 1.645) &= 0.5 - P(0 \le Z \le 1.645) \\ &= 0.5 - 0.45 \\ &= 0.05 \\ P(Z > 2.33) &= 0.5 - P(0 \le Z \le 2.33) \\ &= 0.5 - 0.49 \\ &= 0.01 \end{aligned}$$

Hence for a single-tail test (Right-tail or Left-tail) we compare the computed value of $|Z|$ with 1·645 (at 5% level) and 2·33 (at 1% level) and accept or reject H_0 accordingly.

Important Remark. In the theoretical discussion that follows in the next sections, the samples under consideration are supposed to be large. For practical purposes, sample may be regarded as large if $n > 30$.

12·9. Sampling of Attributes. Here we shall consider sampling from a population which is divided into two mutually exclusive and collectively exhaustive classes-one class possessing a particular attribute, say A, and the other class not possessing that attribute, and then note down the number of persons in the sample of size n, possessing that attribute. The presence of an attribute in sampled unit may be termed as success and its absence as failure. In this case a sample of n observations is identified with that of a series of n independent Bernoulli trials with constant probability P of success for each trial. Then the probability of x successes in n trials, as given by the binomial probability distribution is

$$p(x) = {}^nC_x P^x Q^{n-x} \; ; x = 0, 1, 2, \ldots, n.$$

12·9·1. Test for Single Proportion. If X is the number of successes in n independent trials with constant probability P of success for each trial (*c.f.* § 7·2·1)

$$E(X) = nP \quad \text{and} \quad V(X) = nPQ,$$

where $Q = 1 - P$, is the probability of failure.

It has been proved that for large n, the binomial distribution tends to normal distribution. Hence for large n, $X \sim N(nP, nPQ)$ *i.e.*,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1) \qquad \ldots(12{\cdot}4)$$

and we can apply the normal test.

Remarks 1. In a sample of size n, let X be the number of persons possessing the given attribute. Then

Observed proportion of successes $= X/n = p$, (say).

$$\therefore \quad E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}nP = P$$

$$\Rightarrow \quad E(p) = P \qquad \ldots(12{\cdot}4a)$$

Thus the sample proportion 'p' gives an unbiased estimate of the population proportion P.

Also
$$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}nPQ = \frac{PQ}{n}$$

$$\therefore \quad \text{S.E.}(p) = \sqrt{PQ/n} \qquad \ldots(12{\cdot}4b)$$

Since X and consequently X/n is asymptotically normal for large n, the normal test for the proportion of successes becomes

$$Z = \frac{p - E(p)}{\text{S.E.}(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1) \qquad \ldots(12{\cdot}4c)$$

2. If we have sampling from a finite population of size N, then

$$S.E.(p) = \sqrt{\left(\frac{N - n}{N - 1}\right) \cdot \frac{PQ}{n}} \qquad \ldots(12{\cdot}4d)$$

3. Since the probable limits for a normal variate X are $E(X) \pm 3\sqrt{V(X)}$, the probable limits for the observed proportion of successes are :

$$E(p) \pm 3 \text{ S.E. }(p), \; i.e., \; P \pm 3\sqrt{PQ/n}.$$

If P is not known then taking p (the sample proportion) as an estimate of P, the probable limits for the proportion in the population are :

$$p \pm 3\sqrt{pq/n} \qquad \ldots(12\cdot4e)$$

However, the limits for P at level of significance α are given by :

$$p \pm z_\alpha \sqrt{pq/n}, \qquad \ldots(12\cdot4f)$$

where z_α is the significant value of Z at level of significance α.

In particular 95% confidence limits for P are given by :

$$p \pm 1\cdot96\sqrt{pq/n}, \qquad \ldots(12\cdot4g)$$

and 99% confidence limits for P are given by

$$p \pm 2\cdot58\sqrt{pq/n} \qquad \ldots(12\cdot4h)$$

Example 12·1. *A dice is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the dice cannot be regarded as an unbiased one and find the limts between which the probability of a throw of 3 or 4 lies.*

Solution. If the coming of 3 or 4 is called a success, then in usual notations we are given

$$n = 9{,}000;\ X = \text{Number of successes} = 3{,}240$$

Under the null hypothesis (H_0) *that the dice is an unbiased one*, we get

P = Probability of success = Probability of getting a 3 or 4 $= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

Alternative hypothesis, $H_1 : p \neq \frac{1}{3}$, (*i.e.*, dice is biased).

We have $Z = \dfrac{X - nP}{\sqrt{nQP}} \sim N(0, 1)$, since n is large.

Now $\quad Z = \dfrac{3240 - 9000 \times 1/3}{\sqrt{9000 \times (1/3) \times (2/3)}} = \dfrac{240}{\sqrt{2000}} = \dfrac{240}{44\cdot73} = 5\cdot36$

Since $|Z| > 3$, H_0 is rejected and we conclude that the dice is almost certainly biased.

Since dice is not unbiased, $P \neq \frac{1}{3}$. The probable limits for 'P' are given by :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = p \pm 3\sqrt{pq/n},$$

where $\hat{P} = p = \dfrac{3240}{9000} = 0\cdot36$ and $\hat{Q} = q = 1 - p = 0\cdot64$.

Hence the probable limits for the population proportion of successes may be taken as

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0\cdot36 \pm 3\cdot\sqrt{\frac{0\cdot36 \times 0\cdot64}{9000}} = 0\cdot36 \pm 3 \times \frac{0\cdot6 \times 0\cdot8}{30\sqrt{10}}$$

$$= 0\cdot360 \pm 0\cdot015 = 0\cdot345 \text{ and } 0\cdot375.$$

Hence the probability of getting 3 or 4 almost certainly lies between 0·345 and 0·375.

Example 12·2. *A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the S.E. of the* *proportion of bad ones in a sample of this size is 0·015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8·5 and 17·5.*

Solution. Here we are given $n = 500$

$$X = \text{Number of bad pineapples in the sample} = 65$$

$$p = \text{Proportion of bad pineapples in the sample} = \frac{65}{500} = 0{\cdot}13$$

$$\therefore \quad q = 1 - p = 0{\cdot}87$$

Since P, the proportion of bad pineapples in the consignment is not known, we may take (as in the last example)

$$\hat{P} = p = 0{\cdot}13, \quad \hat{Q} = q = 0{\cdot}87$$

$$\text{S.E. of proportion} = \sqrt{\hat{P}\hat{Q}/n} = \sqrt{0{\cdot}13 \times 0{\cdot}87/500} = 0{\cdot}015$$

Thus, the limits for the proportion of bad pineapples in the consignment are :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0{\cdot}130 \pm 3 \times 0{\cdot}015 = 0{\cdot}130 \pm 0{\cdot}045 = (0{\cdot}085,\ 0{\cdot}175)$$

Hence the percentage of bad pineapples in the consignment lies almost certainly between 8·5 and 17·5.

Example 12·3. *A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain the 98% confidence limits for the percentage number of bad apples in the consignment.*

$$\left[\int_0^{2{\cdot}33} \phi(t)\, dt = 0{\cdot}49 \text{ nearly} \right]$$

Solution. We have :

$$p = \text{Proportion of bad apples in the sample} = \frac{60}{500} = 0{\cdot}12$$

Since the significant value of Z at 98% confidence coefficient (level of significance 2%) is given to be 2·33, 98% confidence limits for population proportion are :

$$p \pm 2{\cdot}33\sqrt{pq/n} = 0{\cdot}12 \pm 2{\cdot}33\sqrt{0{\cdot}12 \times 0{\cdot}88/500}$$

$$= 0{\cdot}12 \pm 2{\cdot}33 \times \sqrt{0{\cdot}0002112} = 0{\cdot}12 \pm 2{\cdot}33 \times 0{\cdot}01453$$

$$= 0{\cdot}12000 \pm 0{\cdot}03385 = (0{\cdot}08615,\ 0{\cdot}15385)$$

Hence 98% confidence limits for percentage of bad apples in the consignment are (8·61, 15·38).

Example 12·4. *In a sample of 1,000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this State at 1% level of significance ?*

Solution. In the usual notations we are given $n = 1{,}000$

$$X = \text{Number of rice eaters} = 540$$

$$\therefore \quad p = \text{Sample proportion of rice eaters} = \frac{X}{n} = \frac{540}{1000} = 0{\cdot}54$$

Null Hypothesis, H_0 : Both rice and wheat are equally popular in the State so that

$$P = \text{Population proportion of rice eaters in Maharashtra} = 0{\cdot}5$$

$$\Rightarrow \quad Q = 1 - P = 0{\cdot}5$$

Alternative Hypothesis, $H_1 : P \neq 0{\cdot}5$ (two-tailed alternative).

Test Statistic. Under H_0, the test statistic is

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since } n \text{ is large).}$$

Now $$Z = \frac{0{\cdot}54 - 0{\cdot}50}{\sqrt{0{\cdot}5 \times 0{\cdot}5/1000}} = \frac{0{\cdot}04}{0{\cdot}0138} = 2{\cdot}532$$

Conclusion. The significant or critical value of Z at 1% level of significance for two-tailed test is 2·58. Since computed $Z = 2{\cdot}532$ is less than 2·58, it is not significant at 1% level of significance. Hence the null hypothesis is accepted and we may conclude that rice and wheat are equally popular in Maharashtra State.

Example 12·5. *Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% level. (Use Large Sample Test.)*

[*Patna Univ. B.Sc. (Hons.), 1992; Bombay Univ. B.Sc. 1987*]

Solution. In the usual notations, we are given $n = 20$.

X = Number of persons who survived after attack by a disease = 18

$$p = \text{Proportion of persons survived in the sample} = \frac{18}{20} = 0{\cdot}90$$

Null Hypothesis, $H_0 : P = 0{\cdot}85$, *i.e.*, the proportion of persons survived after attack by a disease in the lot is 85%.

Alternative Hypothesis, $H_1 : P > 0{\cdot}85$ (Right-tail alternative).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since sample is large).}$$

Now $$Z = \frac{0{\cdot}90 - 0{\cdot}85}{\sqrt{0{\cdot}85 \times 0{\cdot}15/20}} = \frac{0{\cdot}05}{0{\cdot}079} = 0{\cdot}633$$

Conclusion. Since the alternative hypothesis is one-sided (right-tailed), we shall apply right-tailed test for testing significance of Z. The significant value of Z at 5% level of significance for right-tail test is + 1·645. Since computed value of $Z = 0{\cdot}633$ is less than 1·645, it is not significant and we may accept the null hypothesis at 5% level of significance.

12·9·2. Test of Significance for Difference of Proportions. Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say A, among their members. Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two populations respectively. Then sample proportions are given by

$$p_1 = X_1/n_1 \quad \text{and} \quad p_2 = X_2/n_2$$

If P_1 and P_2 are the population proportions, then

$$E(p_1) = P_1,\ E(p_2) = P_2 \qquad [c.f.\ \text{Equation (12·4}a)]$$

and

$$V(p_1) = \frac{P_1Q_1}{n_1} \quad \text{and} \quad V(p_2) = \frac{P_2Q_2}{n_2}$$

Since for large samples, p_1 and p_2 are asymptotically normally distributed, $(p_1 - p_2)$ is also normally dstributed. Then the standard variable corresponding to the difference $(p_1 - p_2)$ is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1)$$

Under the *null hypothesis* $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference between the sample proportions, we have

$$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0 \qquad (\text{Under } H_0)$$

Also $\quad V(p_1 - p_2) = V(p_1) + V(p_2)$,

the covariance term $\text{Cov}(p_1, p_2)$ vanishes, since sample proportions are independent.

$$\therefore \qquad V(p_1 - p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

since under $H_0 : P_1 = P_2 = P$, (say), and $Q_1 = Q_2 = Q$.

Hence under $H_0 : P_1 = P_2$, the test statistic for the difference of proportions becomes

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \ldots(12·5)$$

In general, we do not have any information as to the proportion of A's in the populations from which the samples have been taken. Under $H_0 : P_1 = P_2 = P$, (say), an unbiased estimate of the population proportion P, based on both the samples is given by

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} \qquad \ldots(12·5a)$$

The estimate is unbiased, since

$$E(\hat{P}) = \frac{1}{n_1 + n_2} E\left[n_1p_1 + n_2p_2\right] = \frac{1}{n_1 + n_2}\left[n_1E(p_1) + n_2E(p_2)\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1P_1 + n_2P_2\right] = P \qquad [\because P_1 = P_2 = P, \text{ under } H_0]$$

Thus (12·5) along with (12·5a) gives the required test statistic.

Remarks 1. Suppose we want to test the significance of the difference between p_1 and p, where

$$p = \frac{(n_1p_1 + n_2p_2)}{(n_1 + n_2)}$$

gives a pooled estimate of the population proportion on the basis of both the samples. We have

$$V(p_1 - p) = V(p_1) + V(p) - 2\ \text{Cov}\ (p_1, p) \qquad \ldots(*)$$

Since p_1 and p are not independent, Cov $(p_1, p) \neq 0$.

$$\text{Cov}\ (p_1, p) = E[\{p_1 - E(p_1)\}\{p - E(p)\}]$$

$$= E\left[\{p_1 - E(p_1)\}\ \left\{\frac{1}{n_1 + n_2}\{n_1p_1 + n_2p_2 - E(n_1p_1 + n_2p_2)\}\right\}\right]$$

$$= \frac{1}{n_1 + n_2} E\left[\{p_1 - E(p_1)\}\ \left\{n_1(p_1 - E(p_1)) + n_2(p_2 - E(p_2))\right\}\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1E\left\{p_1 - E(p_1)\right\}^2 + n_2E\left\{(p_1 - E(p_1))(p_2 - E(p_2))\right\}\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1V(p_1) + n_2\ \text{Cov}\ (p_1, p_2)\right]$$

$$= \frac{1}{n_1 + n_2} n_1 V(p_1), \qquad [\because \text{Cov}\ (p_1, p_2) = 0]$$

$$= \frac{n_1}{n_1 + n_2} \cdot \frac{pq}{n_1} = \frac{pq}{n_1 + n_2}$$

$$\text{Also Var}\ (p) = \frac{1}{(n_1 + n_2)^2} E\left[(n_1p_1 + n_2p_2) - E(n_1p_1 + n_2p_2)\right]^2$$

$$= \frac{1}{(n_1 + n_2)^2}\left[n_1^2\ \text{Var}\ (p_1) + n_2^2\ \text{Var}\ (p_2)\right],$$

covariance term vanishes since p_1 and p_2 are independent.

$$\therefore \quad \text{Var}\ (p) = \frac{1}{(n_1 + n_2)^2}\left[n_1^2 \cdot \frac{pq}{n_1} + n_2^2 \cdot \frac{pq}{n_2}\right]$$

$$= \frac{pq}{n_1 + n_2}$$

Substituting in (*) and simplifying, we shall get

$$V(p_1 - p) = \frac{pq}{n_1} + \frac{pq}{n_1 + n_2} - 2\frac{pq}{n_1 + n_2} = pq\left[\frac{n_2}{n_1(n_1 + n_2)}\right]$$

Thus, the test statistic in this case becomes

$$Z = \frac{p_1 - p}{\sqrt{\dfrac{n_2}{(n_1 + n_2)} \cdot \dfrac{pq}{n_1}}} \sim N(0, 1) \qquad \ldots(12\cdot5b)$$

2. Suppose the population proportions P_1 and P_2 are given to be distinctly different, *i.e.*, $P_1 \neq P_2$ and we want to test if the difference $(P_1 - P_2)$ in population proportions is likely to be hidden in simple samples of sizes n_1 and n_2 from the two populations respectively.

We have seen that in the usual notations,

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\text{S.E.}(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0, 1)$$

Here sample proportions are not given. If we set up the *null hypothesis* $H_0 : p_1 = p_2$, *i.e.*, the samples will not reveal the difference in the population proportions or in other words the difference in population proportions is likely to be hidden in sampling, the test statistic becomes

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0, 1) \qquad \ldots(12\cdot5c)$$

Example 12·6. *Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are same against that they are not, at 5% level.* **[*Agra Univ. M.A., 1992*]**

Solution. *Null Hypothesis* $H_0 : P_1 = P_2 = P$, (say), *i.e.*, there is no significant difference between the opinion of men and women as far as proposal of flyover is concerned.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

We are given :

$n_1 = 400$, X_1 = Number of men favouring the proposal = 200

$n_2 = 600$, X_2 = Number of women favouring the proposal = 325

$\therefore \quad p_1$ = Proportion of men favouring the proposal in the sample

$$= \frac{X_1}{n_1} = \frac{200}{400} = 0\cdot5$$

p_2 = Proportion of women favouring the proposal in the sample

$$= \frac{X_2}{n_2} = \frac{325}{600} = 0\cdot541$$

Test Statistic. Since samples are large, the test statistic under the Null Hypothesis, H_0 is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $\hat{P} = \dfrac{n_1p_1 + n_2p_2}{n_1 + n_2} = \dfrac{X_1 + X_2}{n_1 + n_2} = \dfrac{200 + 325}{400 + 600} = \dfrac{525}{1000} = 0\cdot525$

$\Rightarrow \quad \hat{Q} = 1 - \hat{P} = 1 - 0\cdot525 = 0\cdot475$

$$\therefore \quad Z = \frac{0\cdot500 - 0\cdot541}{\sqrt{0\cdot525 \times 0\cdot475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}}$$

$$= \frac{-0.041}{\sqrt{0.525 \times 0.475 \times (10/2{,}400)}}$$

$$= \frac{-0.041}{\sqrt{0.001039}} = \frac{-0.041}{0.0323} = -1.269$$

Conclusion. Since $|Z| = 1.269$ which is less than 1·96, it is not significant at 5% level of significance. Hence H_0 may be accepted at 5% level of significance and we may conclude that men and women do not differ significantly as regards proposal of flyover is concerned.

Example 12·7. *A company has the head office at Calcutta and a branch at Bombay. The personnel director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Calcutta, 62% favoured the new plan. At Bombay out of a sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level ?*

Solution. In the usual notations, we are given :

$$n_1 = 500, p_1 = 0.62 \text{ and } n_2 = 400, p_2 = 1 - 0.41 = 0.59$$

Null hypothesis, $H_0 : P_1 = P_2$, *i.e.,* there is no significant difference between the two groups in their attitude towards the new plan.

Alternative hypothesis, $H_1 : P_1 \neq P_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{p_1 - p_2}{\text{S.E. }(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{500 \times 0.62 + 400 \times 0.59}{500 + 400} = 0.607$$

and $$\hat{Q} = 1 - \hat{P} = 0.393$$

$\therefore$ $$Z = \frac{0.62 - 0.59}{\sqrt{0.607 \times 0.393 \times \left(\frac{1}{500} + \frac{1}{400}\right)}}$$

$$= \frac{0.03}{\sqrt{0.00107}} = \frac{0.03}{0.0327} = 0.917.$$

Critical region. At 5% level of significance, the critical value of Z for a two-tailed test is 1·96. Thus the critical region consists of all values of $Z \geq 1.96$ or $Z \leq -1.96$.

Conclusion. Since the calculated value of $|Z| = 0.917$ is less than the critical value of Z (1·96), it is not significant at 5% level of significance. Hence the data do not provide us any evidence against the null hypothesis which may be accepted, and we conclude that there is no significant difference between the two groups in their attitude towards the new plan.

Example 12·8. *Before an increase in excise duty on tea, 800 persons out of a sample of 1,000 persons were found to be tea drinkers. After an increase in* *duty, 800 people were tea drinkers in a sample of 1,200 people. Using standard error of proportion, state whether there is a significant decrease in the consumption of tea after the increase in excise duty ?*

Solution. In the usual notations, we have $n_1 = 1{,}000$; $n_2 = 1{,}200$

p_1 = Sample proportion of tea drinkers before increase in excise duty

$$= \frac{800}{1000} = 0{\cdot}80$$

p_2 = Sample proportion of tea drinkers after increase in excise duty

$$= \frac{800}{1200} = 0{\cdot}67$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.,* there is no significant difference in the consumption of tea before and after the increase in excise duty.

Alternative Hypothesis, $H_1 : P_1 > P_2$ (Right-tailed alternative).

Test Statistic. Under the null hypothesis, the test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \text{(Since samples are large)}$$

where

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{16}{22}, \text{ and } \hat{Q} = 1 - \hat{P} = \frac{6}{22}$$

$$\therefore \quad Z = \frac{0{\cdot}80 - 0{\cdot}67}{\sqrt{\frac{16}{22} \times \frac{6}{22} \times \left(\frac{1}{1000} + \frac{1}{1200}\right)}}$$

$$= \frac{0{\cdot}13}{\sqrt{\frac{16}{22} \times \frac{6}{22} \times \frac{11}{6000}}} = \frac{0{\cdot}13}{0{\cdot}019} = 6{\cdot}842$$

Conclusion. Since Z s much greater than 1·645 as well as 2·33 (since test is one-tailed), it is highly significant at both 5% and 1% levels of significance. Hence,

we reject the null hypothesis H_0 and conclude that there is a significant decease in the consumption of tea after increase in the excise duty.

Example 12·9. *A cigarette manufacturing firm claims that its brand A of the cigarettes outsells its brand B by 8%. If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another random sample of 100 smokers prefer brand B, test whether the 8% difference is a valid claim. (Use 5% level of significance.)*

Solution. We are given.:

$$n_1 = 200, X_1 = 42 \quad \Rightarrow \quad p_1 = \frac{X_1}{n_1} = \frac{42}{200} = 0{\cdot}21$$

$$n_2 = 100, X_2 = 18 \quad \Rightarrow \quad p_2 = \frac{X_2}{n_2} = \frac{18}{100} = 0{\cdot}18$$

We set up the *Null Hypothesis* that 8% difference in the sale of two brands of cigarettes is a valid claim, *i.e.,* $H_0 : P_1 - P_2 = 0{\cdot}08$.

Alternative Hypothesis : $H_1 : P_1 - P_2 \neq 0{\cdot}08$ (Two-tailed).

Under H_0, the test statistic is (since samples are large)

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{42 + 18}{200 + 100} = \frac{60}{300} = 0{\cdot}20 \Rightarrow \hat{Q} = 1 - \hat{P} = 0{\cdot}80$

$$\therefore \quad Z = \frac{(0{\cdot}21 - 0{\cdot}18) - (0{\cdot}08)}{\sqrt{0{\cdot}2 \times 0{\cdot}8\left(\frac{1}{200} + \frac{1}{100}\right)}} = \frac{-0{\cdot}05}{\sqrt{0{\cdot}16 \times 0{\cdot}015}}$$

$$= \frac{-0{\cdot}05}{\sqrt{0{\cdot}0024}} = \frac{-0{\cdot}05}{0{\cdot}04899} = -1{\cdot}02$$

Since $|Z| = 1{\cdot}02 < 1{\cdot}96$, it is not significant at 5% level of significance. Hence null hypothesis may be retained at 5% level of significance and we may conclude that a difference of 8% in the sale of two brands of cigarettes is a valid claim by the firm.

Example 12·10. *On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30 per cent and the remaining 70 per cent. Consider the first question of this examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is no good at discriminating ability of the type being examined here?*

Solution. Here, we have

n = Total number of candidates = 200

n_1 = The number of candidates in the upper 30% group

$= \frac{30}{100} \times 200 = 60$

n_2 = The number of candidates in the remaining 70% group

$= \frac{70}{100} \times 200 = 140$

X_1 = The number of candidates, with correct answer in the first group = 40

X_2 = The number of candidates, with correct answer in the second group = 80

$\therefore \quad p_1 = \frac{X_1}{n_1} = \frac{40}{60} = 0{\cdot}6666$ and $p_2 = \frac{X_2}{n_2} = \frac{80}{140} = 0{\cdot}5714$

Null Hypothesis, H_0 : There is no significant difference in the sample proportions, *i.e.*, $P_1 = P_2$, *i.e.*, the first question is no good at dicriminating the ability of the type being examined here.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$.

Test Statistic. Under H_0 the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \quad \text{(since samples are large).}$$

where

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{40 + 80}{60 + 140} = 0·6, \quad \hat{Q} = 1 - \hat{P} = 0·4$$

$$\therefore \qquad Z = \frac{0·6666 - 0·5714}{\sqrt{0·6 \times 0·4 \left(\frac{1}{60} + \frac{1}{140}\right)}} = \frac{0·0953}{0·0756} = 1·258$$

Conclusion. Since $|Z| < 1·96$, the data are consistent with the null hypothesis at 5% level of significance. Hence we conclude that the first question is not good enough to distinguish between the ability of the two groups of candidates.

Example 12·11. *In a year there are 956 births in a town A, of which 52·5% were males, while in towns A and B combined, this proportion in a total of 1,406 births was 0·496. Is there any significant difference in the proportion of male births in the two towns ?*

Solution. We are given

$$n_1 = 956, n_1 + n_2 = 1{,}406 \text{ or } n_2 = 1{,}406 - 956 = 450$$

$$p_1 = \text{Proportion of males in the sample of town } A = 0·525.$$

Let p_2 be the proportion of males in the sample (of size n_2) of town B. Then

$$\hat{P} = \text{Proportion of males in both the samples combined.}$$

$$= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0·496 \qquad \text{(Given)}$$

$$\therefore \qquad \frac{956 \times 0·525 + 450 \times p_2}{1{,}406} = 0·496$$

$$\Rightarrow \qquad p_2 = 0·434 \qquad \text{(On simplification)}$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference in the proportion of male births in the two towns A and B.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \text{(Since samples are large)}$$

where $\hat{P} = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0·496, \quad \hat{Q} = 1 - \hat{P} = 0·504$

$$\therefore \qquad Z = \frac{0·525 - 0·434}{\sqrt{0·496 \times 0·504 \left(\frac{1}{956} + \frac{1}{450}\right)}} = \frac{0·091}{0·027} = 3·368$$

Conclusion. Since $|Z| > 3$, the null hypothesis is rejected, *i.e.*, the data are inconsistent with the hypothesis $P_1 = P_2$ and we conclude that there is significant difference in the proportion of male births in the towns A and B.

Example 12·12. *In two large populations, there are 30 and 25 per cent respectively of blue-eyed people. Is this difference likely to be hidden in samples of 1,200 and 900 respectively from the two populations ?*

[*Delhi Univ. B.Sc., 1992*]

Solution. Here, we are given $n_1 = 1200,\ n_2 = 900$.

P_1 = Proportion of blue-eyed people in the first population = 30% = 0·30.

P_2 = Proportion of blue-eyed people in the second population = 25% = 0·25.

$$\therefore \qquad Q_1 = 1 - P_1 = 0{\cdot}70 \text{ and } Q_2 = 1 - P_2 = 0{\cdot}75$$

We set up the *null hypothesis* H_0 that $p_1 = p_2$, *i.e.,* the sample proportions are equal, *i.e.,* the difference in population proportions is likely to be hidden in sampling.

Test Statistic. Under $H_0 : p_1 = p_2$, the test statistic is :

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\dfrac{P_1Q_1}{n_1} + \dfrac{P_2Q_2}{n_2}}} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

$$\therefore \qquad |Z| = \frac{0{\cdot}30 - 0{\cdot}25}{\sqrt{\dfrac{0{\cdot}3 \times 0{\cdot}7}{1{,}200} + \dfrac{0{\cdot}25 \times 0{\cdot}75}{900}}} = \frac{0{\cdot}05}{0{\cdot}0195} = 2{\cdot}56$$

Conclusion. Since $|Z| > 1{\cdot}96$, the null hypothesis $(p_1 = p_2)$, is refuted at 5% level of significance and we conclude that the difference in population proportions is unlikely to be hidden in sampling. In other words, these samples will reveal the difference in the population proportions.

Example 12·13. *In a random sample of 400 students of the university teaching departments, it was found that 300 students failed in the examination. In another random sample of 500 students of the affiliated colleges, the number of failures in the same examination was found to be 300. Find out whether the proportion of failures in the university teaching departments is significantly greater than the proportion of failures in the university teaching departments and affiliated colleges taken together.*

Solution. Here we are given : $n_1 = 400,\ n_2 = 500$

$$p_1 = \frac{300}{400} = 0{\cdot}75,\quad p_2 = \frac{300}{500} = 0{\cdot}60$$

$$\therefore \qquad q_1 = 1 - p_1 = 1 - 0{\cdot}75 = 0{\cdot}25 \text{ and } q_2 = 1 - p_2 = 0{\cdot}40$$

Here we set up the *null hypothesis* H_0 that p_1 and $\hat{p}$, where $\hat{p}$ is the pooled estimate, *i.e.,* proportion of failures in the university teaching departments and affiliated colleges taken together, do not differ significantly.

$$\text{S.E. of } (\hat{p} - p_1) = \sqrt{\frac{\hat{p}\,\hat{q}}{n_1 + n_2} \times \frac{n_2}{n_1}} \qquad [c.f. \text{ (12·5}b\text{) page 12·18}]$$

$$\text{where} \qquad \hat{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{400 \times 0{\cdot}75 + 500 \times 0{\cdot}60}{400 + 500} = 0{\cdot}67$$

$$\hat{q} \doteq 1 - 0{\cdot}67 = 0{\cdot}33$$

$$\therefore \quad \text{S.E. of } (\hat{p} - p_1) = \sqrt{\frac{0{\cdot}67 \times 0{\cdot}33}{400 + 500} \times \frac{500}{400}} = 0{\cdot}018$$

Test Statistic. Under the null hypothesis H_0, the test statistic is :

$$Z = \frac{\hat{p} - p_1}{\text{S.E. of } (\hat{p} - p_1)} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

$$Z = \frac{0{\cdot}67 - 0{\cdot}33}{0{\cdot}018} = \frac{0{\cdot}15}{0{\cdot}018} = 8{\cdot}3$$

Conclusion. Since the calculated value of Z is much greater than 3, it is highly signifcant. Hence null hypothesis H_0 is rejected and we conclude that there is significant difference between p_1 and $\hat{p}$.

Example 12·14. *If for one-half of n events, the chance of success is p and the chance of failure is q, while for the other half the chance of success is q and the chance of failure is p, show that the standard deviation of the number of successes is the same as if the chance of successes were p in all the cases, i.e., $\sqrt{npq}$ but that the mean of the number of successes is n/2 and not np.*

Solution. Let X_1 and X_2 denote the number of successes in the first half and the second half of n events respectively. Then according to the given conditions, we have

$$\left.\begin{aligned} E(X_1) &= \frac{n}{2}p \\ V(X_1) &= \frac{n}{2}pq \end{aligned}\right\} \text{and} \quad \left.\begin{aligned} E(X_2) &= \frac{n}{2}q \\ V(X_2) &= \frac{n}{2}pq \end{aligned}\right\}$$

The mean and variance of the number of successes in all the n events are given by $\quad E(X_1 + X_2) = E(X_1) + E(X_2) = \frac{n}{2}p + \frac{n}{2}q = \frac{n}{2}$

and $\quad V(X_1 + X_2) = V(X_1) + (X_2) = \frac{n}{2}pq + \frac{n}{2}qp = npq,$

since the first and second half of events are independent.

Hence the variance is the same as if the probability of success in all the n events is p.

EXERCISE 12(a)

1. (*a*) There are 2 populations and P_1 and P_2 are the proportion of members in the two populations belonging to 'low-income' group. It is disired to test the hypothesis $H_0 : P_1 = P_2$. Explain clearly, the procedure that you would follow to carry out the above test at 5% level of significance.

State the theorem on which the above test is based.

In respect of the above 2 populations, if it is claimed that P_1, the proportion of 'low-income' group in the first population is greater than P_2, how will you modify the procedure to test this claim (at 5% level) ?

(*b*) Take a concrete illustration and in relation to this illustration, explain the following terms :—

(*i*) Null hypothesis and alternative hypothesis.

(*ii*) Type I and Type II errors.

(*iii*) Critical Region.

(*c*) Suggest a possible source of bias in the following :

(*i*) The mean income per family in a certain town is sought to be estimated by sampling from motor owners.

(*ii*) Readers of newspapers are sampled by printing in it an invitation to them to send up their observations on some typical event.

(*iii*) A barrel of apples is sampled by taking a handful from the top.

(*iv*) A set of digits is taken by opening a telephone directory at random and choosing the telephone numbers in the order in which they appear on the page.

2. (*a*) Explain clearly the terms "Standard Error" and "Sampling Distribution." Show that in a series of n independent trials with constant probability p of success, the standard error of the proportion of successes is $\sqrt{pq/n}$, where $q = 1 - p$.

(*b*) n individuals fall into one or the other two categories with probabilities p and q $(=1 - p)$, the number in the two categories are x_1 and x_2 $(x_1 + x_2 = n)$. Show that covariance between x_1 and x_2 is $-npq$. Hence obtain the variance of the difference $\left(\frac{x_1}{n} - \frac{x_2}{n}\right)$, between the proportions.

(*c*) Explain clearly the procedure generally followed in testing of a hypothesis. Point out the difference between one-tail and two-tail tests.

(*d*) What do you mean by interval estimation and how would you set up the confidence limits for a parameter from a sample ? Give the formula for 95% confidence limits for mean and proportion. What modifcations do you have to make if the sampling is done from finite population, (*i*) without replacement, (*ii*) with replacement ? [*Calcutta Univ. B.A. (Maths Hons.), 1988*]

3. P_1 and P_2 are the (unknown) proportions of students wearing glasses in two universities A and B. To compare P_1 and P_2, samples of sizes n_1 and n_2 are taken from the two populations and the number of students wearing glasses is found to be x_1 and x_2 respectively. Suggest an unbiased estimate of $(P_1 - P_2)$ and obtain its sampling distribution when n_1 and n_2 are large. Hence explain how to test the hypothesis that $P_1 = P_2$.

4. (*a*) A coin is tossed 10,000 times and it turns up head 5,195 times. Discuss whether the coin may be regarded as unbiased one, explaining briefly the theoretical principles you would use for this purpose. (**Ans.** No.)

(*b*) A biased coin was thrown 400 times and head resulted 240 times. Find the standard error of the observed proportion of heads and deduce that the probability of getting a head in a single throw of the coin lies almost certainly between 0·53 and 0·67. (**Ans.** 0·02445).

(*c*) Experience has shown that 20% of a manufactured product is of the top quality. In one day's production of 400 articles only 50 are of top quality. Show that either the production of the day taken was not a representative sample or the hypothesis of 20% was wrong. (**Ans.** $Z = 3{\cdot}75$)

5. (*a*) In a large consignment of oranges a random sample of 64 oranges revealed that 14 oranges were bad. Is it reasonable to assume that 20% of the oranges were bad?

(*b*) By a mobile court checking in certain buses it was found that out of 1000 people checked on a certain day at Red Fort, 10 persons were found to be ticketless travellers. If daily 1 lakh passengers travel by the buses, find out the estimated limits to the ticketless travellers. (**Ans.** 997 to 1003)

(*c*) In a random sample of 81 items taken from a large consignment some were found to be defective. If the standard error of the proportion of defective items in the sample is 1/18, find 95% confidence limits of the percentage of defective items in the consignment.

[***Madras Univ. B.Sc. (Stat. Main), 1991***]

6. (*a*) In some dice throwing experiments Weldon threw dice 75,145 times and of these 49,152 yielded a 4, 5 or 6. Is this consistent with the hypothesis that the dice was unbiased?

Hint. H_0 : Dice is unbiased, *i.e.*, $P = \frac{3}{6} = \frac{1}{2} = 0{\cdot}5$; $H_1 : P \neq \frac{1}{2}$

Test Statistic. Under H_0, $Z = \dfrac{p-P}{\sqrt{PQ/n}} = \dfrac{0{\cdot}654 - 0{\cdot}5}{\sqrt{0{\cdot}5 \times 0{\cdot}5/75145}} = \dfrac{0{\cdot}154}{0{\cdot}0018}$

Ans. No.

(*b*) 1,000 apples are taken from a large consignment and 100 are found to be bad. Estimate the percentage of bad apples in the consignment and assign the limits within which the percentage lies.

7. (*a*) A personnel manager claims that 80 per cent of all single women hired for secretarial job get married and quit work within two years after they are hired. Test this hypothesis at 5% level of significance if among 200 such secretaries, 112 got married within two years after they were hired and quit their jobs.

(*b*) A manufacturer claimed that at least 98% of the steel pipes which he supplied to a factory conformed to specifications. An examination of a sample of 500 pieces of pipes revealed that 30 were defective. Test this claim at a significance level of (*i*) 0·05, (*ii*) 0·01.

Hint. X = No. of pipes conforming to specifications in the sample.
$= 500 - 30 = 470$

p = Sample proportion of pipes conforming to specifications
$= \dfrac{470}{500} = 0{\cdot}94$

$H_0 : P = 0{\cdot}98$, *i.e.*, the proportion of pipes conforming to specifications in the lot is 98%.

$H_1 : P < 0{\cdot}98$ (Left-tail alternative)

Test Statistic. $Z = \dfrac{p-P}{\sqrt{PQ/n}} = \dfrac{0{\cdot}94 - 0{\cdot}98}{\sqrt{0{\cdot}98 \times 0{\cdot}02/500}}$

(*c*) A social worker believes that fewer than 25% of the couples in a certain area ever used any form of birth control. A random sample of 120 couples was contacted. Twenty of them said they had used some method of birth control. Comment on the social worker's belief.

$H_0 : P = 0{\cdot}25, H_1 : P < 0{\cdot}25$ (left-Tailed)

8. In a random sample of 800 adults from the population of a certain large city, 600 are found to have dark hair. In a random sample of 1,000 adults from the habitants of another large city, 700 are dark haired. Show that the difference of the proportion of dark haired people is nearly 2·4 times the standard error of the difference for samples of above sizes.

9. (*a*) In a random sample of 100 men taken from village A, 60 were found to be consuming alcohol. In another sample of 200 men taken form village B, 100 were found to be consuming alcohol. Do the two villages differ significantly in respect of the proportion of men who consume alcohol ?

[*Delhi Univ. M.A. (Business Eco.), 1987*]

(*b*) In a random sample of 500 men from a particular district of U.P., 300 are found to be smokers. In one of 1,000 men from another district, 550 are smokers. Do the data indicate that the two districts are significantly different with respect to the prevalence of smoking among men ?

Ans. $Z = 1{\cdot}85$, (not significant). (*Delhi Univ. B.Sc., 1991*)

10. A company is considering two different television advertisements for promotion of a new product. Management believed that the advertisement A is more effective than advertisement B. Two test market areas with virtually identical consumer characteristics are selected; A is used in one area and B in other area. In a random sample of 60 customers who saw A, 18 tried the product. In another random sample of 100 customers who saw B, 22 tried the product. Does this indicate that advertisement A is more effective than advertisement B, if a 5% level of significance is used ? Given critical value at 5% level is 1·96 and at 10% level of significance is 1·645.

[*Delhi Univ. M.C.A., 1990*]

11. (*a*) 1,000 apples kept under one type of storage were found to show rotting to the extent of 4%. 1,500 apples kept under another kind of storage showed 3% rotting. Can it be reasonably concluded that the second type of storage is superior to the first ?

(*b*) In a referendum submitted to the students body at a university, 850 men and 566 women voted. 530 of the men and 304 of the women voted yes. Does this indicate a significant difference of opinion on the matter at 1% level, between men and women students. [**Ans.** $Z = 3{\cdot}2$, (significant).]

(*c*) In a simple sample of 600 high school students from a State, 400 are found to use dot pens. In one of 900 from a neighbouring State, 450 are found to use dot pens. Do the data indicate that the States are significantly different with respect to the habit of using dot pens among the students ? (**Ans.** Yes.)

12. (*a*) A firm, manufacturing dresses for children, sent out advertisement through mail. Two groups of 1,000 each were contacted; the first group having been contacted in white covers while the second in blue covers. 20% from the first while 28% from the second replied.

Do you think that blue envelopes help the sales ?

(*b*) A machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved ?

Hint. We are given : $n_1 = 500$, and $n_2 = 100$

$$p_1 = \frac{16}{500} = 0{\cdot}032; \quad p_2 = \frac{3}{100} = 0{\cdot}030$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference in the machine before overhauling and after overhauling. In other words, the machine has not improved after overhauling.

Alternative Hypothesis, $H_1 : P_2 < P_1$ or $P_1 > P_2$.

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{16 + 3}{500 + 100} = \frac{19}{600} = 0{\cdot}032$$

$$\text{S.E. } (p_1 - p_2) = \sqrt{0{\cdot}032 \times 0{\cdot}968 \left(\frac{1}{500} + \frac{1}{100}\right)} = 0{\cdot}0193$$

$$Z = \frac{0{\cdot}032 - 0{\cdot}030}{0{\cdot}0193} = \frac{0{\cdot}002}{0{\cdot}0193} = 1{\cdot}04$$

Since $Z < 1{\cdot}645$ (Right-tailed test), it is not significant at 5% level of significance.

(*c*) In a large city *A*, 25% of a random sample of 900 school boys had defective eye-sight. In another large city *B*, 15·5% of a random sample of 1,600 school boys had the same defect. Is this difference between the two proportions significant ? (**Ans.** Not significant.)

13. (*a*) A candidate for election made a speech in city *A* but not in *B*. A sample of 500 voters from city *A* showed that 59·6% of the voters were in favour of him, whereas a sample of 300 voters from city *B* showed that 50% of the voters favoured him. Discuss whether his speech could produce any effect on voters in city *A*. Use 5% level.

Ans. $|Z| = 2{\cdot}67$. Yes.

(*b*) In a large city, 16 out of a random sample of 500 men were found to be drinkers. After the heavy increase in tax on intoxicants another random sample of 100 men in the same city included 3 drinkers. Was the observed decrease in the proportion of drinkers significant after tax increase ?

Ans. $H_0 : P_1 = P_2, H_1 : P_1 > P_2$; $Z = 1{\cdot}04$. Not sigificant.

14. The sex ratio at birth is sometimes given by the ratio of male to female births instead of the proportion of male to total births. If z is the ratio, *i.e.*, $z = p/q$, show that the standard error of z is approximately $\frac{1}{1+z}\sqrt{\left(\frac{z}{n}\right)}$

n being large, so that deviations are small compared with mean.

12·10. Sampling of Variables. In the present section we will discuss in detail the sampling of variables such as height, weight, age, income, etc. In the case of sampling of variables each member of the population provides the value of the variable and the aggregate of these values forms the frequency distribution of the population. From the population, a random sample of size n

can be drawn by any of the sampling methods discussed before which is same as choosing n values of the given variable from the distribution.

12·11. Unbiased Estimate for population Mean (μ) and Variance (σ^2). Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a large population $X_1, X_2, \ldots X_N$ (of size N) with mean μ and variance σ^2. Then the sample mean ($\bar{x}$) and variance (s^2) are given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{and} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Now $$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i)$$

Since x_i is a sample observation from the population X_i, ($i = 1, 2, \ldots, N$) it can take any one of the values $X_1, X_2, \ldots, X_N$ each with equal probability $1/N$.

$$\therefore \quad E(x_i) = \frac{1}{N}X_1 + \frac{1}{N}X_2 + \ldots + \frac{1}{N}X_N$$

$$= \frac{1}{N}(X_1 + X_2 + \ldots + X_N) = \mu \qquad \ldots(1)$$

$$\therefore \quad E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(\mu) = \frac{1}{n}n\mu \quad \Rightarrow \quad E(\bar{x}) = \mu \qquad \ldots(12\cdot6)$$

Thus the sample mean ($\bar{x}$) is an unbiased estimate of the population mean (μ).

Now $$E(s^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(x_i^2) - E(\bar{x})^2 \qquad \ldots(2)$$

We have $$V(x_i) = E[x_i - E(x_i)]^2 = E(x_i - \mu)^2, \qquad \text{[From (1)]}$$

$$= \frac{1}{N}\left[(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_N - \mu)^2\right] = \sigma^2 \qquad \ldots(3)$$

Also we know that

$$V(x) = E(x^2) - [E(x)]^2 \Rightarrow E(x^2) = V(x) + \{E(x)\}^2 \qquad \ldots(4)$$

In particular,

$$E(x_i^2) = V(x_i) + \{E(x_i)\}^2 = \sigma^2 + \mu^2 \qquad \ldots(5)$$

Also from (4), $E(\bar{x}^2) = V(\bar{x}) + \{E(\bar{x})\}^2$

But $V(\bar{x}) = \dfrac{\sigma^2}{n}$, where σ^2 is the population variance. [c.f. § 12·13]

$$\therefore \quad E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2 \qquad \text{[Using (12·6)]} \qquad \ldots(5a)$$

Substituting from (5) and (5a) in (2) we get

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right)$$

$$= \frac{1}{n}n(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right) = \left(1-\frac{1}{n}\right)\sigma^2$$

$$= \frac{n-1}{n}\sigma^2 \qquad \ldots(12\cdot7)$$

Since $E(s^2) \neq \sigma^2$, sample variance is not an unbiased estimate of population variance.

From (12·7), we get

$$\frac{n}{n-1}E(s^2) = \sigma^2 \Rightarrow E\left(\frac{ns^2}{n-1}\right) = \sigma^2$$

$$\Rightarrow \quad E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \sigma^2 \text{ i.e., } E(S^2) = \sigma^2 \qquad \ldots(12\cdot8)$$

where $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad \ldots(12\cdot8a)$$

$\therefore$ S^2 is an unbiased estimate of the population variance σ^2.

Aliter for $E(s^2)$.

$$s^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{1}{n}\left[\sum_{i=1}^{n}\left\{(x_i-\mu)-(\bar{x}-\mu)\right\}^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + n(\bar{x}-\mu)^2 - 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)\right]$$

But $\sum_i (x_i-\mu) = \sum_i x_i - n\mu = n\bar{x} - n\mu = n(\bar{x}-\mu)$

$$\therefore \quad s^2 = \frac{1}{n}\left\{\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\bar{x}-\mu)^2$$

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{n}E(x_i-\mu)^2 - E(\bar{x}-\mu)^2$$

$$\therefore \quad = \frac{1}{n}\sum_{i=1}^{n}E\{x_i-E(x_i)\}^2 - E\{\bar{x}-E(\bar{x})\}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}V(x_i) - V(\bar{x}) = \left(1-\frac{1}{n}\right)\sigma^2$$

Remarks 1. Here we see that although sample mean is an unbiased estimate of population mean, sample variance is not an unbiased estimate of population variance. However, an unbiased estimate of of σ^2 is given by S^2, given in equation (12·8*a*).

S^2 plays a very important role in sampling theory, particularly in small sampling theory. Whenever σ^2 is not known, its estimate S^2 given by (12·8*a*) is used for practical purposes.

2. We have $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

$$\Rightarrow \qquad ns^2 = (n-1)S^2 \qquad \therefore \quad s^2 = \left(1 - \frac{1}{n}\right) S^2$$

Hence for large samples *i.e.*, for $n \to \infty$, we have $s^2 \to S^2$. In other words, for large samples (*i.e.*, $n \to \infty$), we may take

$$\hat{\sigma}^2 = s^2 \qquad \qquad ...(12·8b)$$

12·12. Standard Error of Sample Mean. *The variance of the sample mean is σ^2/n, where σ is the population standard deviation and n is the size of the random sample.*

The S.E. of mean of a random sample of size n from a population with variance σ^2 is $\sigma/\sqrt{n}$.

Proof. Let x_i, $(i = 1, 2, ..., n)$ be a random sample of size n from a population with variance σ^2, then the sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n)$$

$$\therefore \quad V(\bar{x}) = V\left[\frac{1}{n}(x_1 + x_2 + ... + x_n)\right] = \frac{1}{n^2} V(x_1 + x_2 + ... + x_n)$$

$$= \frac{1}{n^2}\left[V(x_1) + V(x_2) + ... + V(x_n)\right],$$

the covariance terms vanish since the sample observations are independent, [*c.f.* Remark (*ii*) § 6·6]

But $V(x_i) = \sigma^2, (i = 1, 2, ..., n)$ [From (3) of § 12·11]

$$\therefore \quad V(\bar{x}) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

$$\Rightarrow \qquad \text{S.E.}(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \qquad \qquad ...(12·9)$$

12·13. Test of Significance for Single Mean. We have proved that if x_i, $(i = 1, 2, ..., n)$ is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance σ^2/n, *i.e.*, $\bar{x} \sim N(\mu, \sigma^2/n)$. However, this result holds, *i.e.*, $\bar{x} \sim N(\mu, \sigma^2/n)$, even in random sampling from non-normal population provided the sample size n is large [*c.f.* Central Limit Theorem, § 8·10].

Thus for large samples, the *standard normal variate* corresponding to $\bar{x}$ is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Under the *null hypothesis*, H_0 that the sample has been drawn from a population with mean μ and variance σ^2, *i.e.*, there is no significant difference between the sample mean ($\bar{x}$) and population mean (μ), the test statistic (for large samples), is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \text{...(12·9}a\text{)}$$

Remarks 1. If the population s.d. σ is unknown then we use its estimate provided by the sample variance given by [See (12·8*b*)]:

$$\hat{\sigma}^2 = s^2 \quad \Rightarrow \quad \hat{\sigma} = s \text{ (for large samples).}$$

2. Confidence limits for μ. 95% confidence interval for μ is given by :

$$|Z| \le 1·96, \text{ i.e., } \left| \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right| \le 1·96$$

$$\Rightarrow \qquad \bar{x} - 1·96\sigma/\sqrt{n} \le \mu \le \bar{x} + 1·96\sigma/\sqrt{n} \qquad \text{...(12·10)}$$

and $\bar{x} \pm 1·96\sigma/\sqrt{n}$ are known as 95% confidence limits for μ. Similarly, 99% confidence limits for μ are $\bar{x} \pm 2·58\sigma/\sqrt{n}$ and 98% confidence limits for μ are $\bar{x} \pm 2·33\sigma/\sqrt{n}$.

However, in sampling from a finite population of size N, the corresponding 95% and 99% confidence limits for μ are respectively

$$\bar{x} \pm 1·96 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \text{ and } \bar{x} \pm 2·58 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \qquad \text{...(12·10}a\text{)}$$

3. The confidence limits for any parameter (P, μ, etc.) are also known as its *fiducial limits*.

Example 12·15. *A sample of 900 members has a mean 3·4 cms,. and s.d. 2·61 cms. Is the sample from a large population of mean 3·25 cms. and s.d. 2·61 cms. ?*

If the population is normal and its mean is unknown, find the 95% and 98% fiducial limits of true mean.

Solution. *Null hypothesis, (H_0)* : The sample has been drawn from the population with mean $\mu = 3·25$ cms., and S.D. $\sigma = 2·61$ cms.

Alternative Hypothesis, $H_1 : \mu \ne 3·25$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ (since } n \text{ is large)}$$

Here, we are given

$$\bar{x} = 3{\cdot}4 \text{ cms.}, n = 900 \text{ cms.}, \mu = 3{\cdot}25 \text{ cms. and } \sigma = 2{\cdot}61 \text{ cms.}$$

$$Z = \frac{3{\cdot}40 - 3{\cdot}25}{2{\cdot}61/\sqrt{900}} = \frac{0{\cdot}15 \times 30}{2{\cdot}61} = 1{\cdot}73$$

Since $|Z| < 1{\cdot}96$, we conclude that the data don't provide us any evidence against the null hypothesis (H_0) which may, therefore, be accepted at 5% level of significance.

95% fiducial limits for the population mean μ are :

$$\bar{x} \pm 1{\cdot}96\, \sigma/\sqrt{n} \Rightarrow 3{\cdot}40 \pm 1{\cdot}96 \times 2{\cdot}61/\sqrt{900}$$

$$\Rightarrow \quad 3{\cdot}40 \pm 0{\cdot}1705, \text{ i.e., } 3{\cdot}5705 \text{ and } 3{\cdot}2295$$

98% fiducial limits for μ are given by :

$$\bar{x} \pm 2{\cdot}33 \frac{\sigma}{\sqrt{n}}, \text{ i.e., } 3{\cdot}40 \pm 2{\cdot}33 \times \frac{2{\cdot}61}{30}$$

$$\Rightarrow \quad 3{\cdot}40 \pm 0{\cdot}2027 \text{ i.e., } 3{\cdot}6027 \text{ and } 3{\cdot}1973$$

Remark. 2·33 is the value z_1 of Z from standard normal probability integrals, such that $P(|Z| > z_1) = 0{\cdot}98 \Rightarrow P(Z > z_1) = 0{\cdot}49$.

Example 12·16. *An insurance agent has claimed that the average age of policyholders who insure through him is less than the average for all agents, which is 30·5 years.*

A random sample of 100 policyholders who had insured through him gave the following age distribution :

Age last birthday	*No. of persons*
16—20	*12*
21—25	*22*
26—30	*20*
31—35	*30*
36—40	*16*

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at the 5% level of significance. You are given that Z (1·645) = 0·95.

Solution. *Null Hypothesis,* $H_0 : \mu = 30{\cdot}5$ years, *i.e.,* the sample mean ($\bar{x}$) and population mean (μ) do not differ significantly.

Alternative Hypothesis, $H_1 : \mu < 30{\cdot}5$ years (Left-tailed alternative).

CALCULATIONS FOR SAMPLE MEAN AND S.D.

Age last birthday	*No. of persons (f)*	*Mid-point x*	$d = \frac{x-28}{5}$	*fd*	*fd*^{2}
16—20	12	18	−2	−24	48
21—25	22	23	−1	−22	22
26—30	20	28	0	0	0
31—35	30	33	1	30	30
36—40	16	38	2	32	64
Total	$N = 100$			$\Sigma fd = 16$	$\Sigma fd^2 = 164$

$$\bar{x} = 28 + \frac{5 \times 16}{100} = 28{\cdot}8 \text{ years} \quad s = 5 \times \sqrt{\frac{164}{100} - \left(\frac{16}{100}\right)^2} = 6{\cdot}35 \text{ years}$$

Since the sample is large, $\hat{\sigma} \simeq s = 6{\cdot}35$ years.

Test Statistic. Under H_0, the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim N(0, 1), \text{ (since sample is large).}$$

Now $$Z = \frac{28{\cdot}8 - 30{\cdot}5}{6{\cdot}35/\sqrt{100}} = \frac{-1{\cdot}7}{0{\cdot}635} = -2{\cdot}681$$

Conclusion. Since computed value of $Z = -2{\cdot}681 < -1{\cdot}645$ or $|Z| = 2{\cdot}681 > 1{\cdot}645$, it is significant at 5% level of significance. Hence we reject the null hypothesis H_0 (Accept H_1) at 5% level of significance and conclude that the insurance agent's claim that the average age of policyholders who insure through him is less than the average for all agents, is valid.

Example 12·17. *As an application of Central Limit Theorem, show that if E is such that $P(|\bar{X} - \mu| < E) > 0{\cdot}95$, then the minimum sample size n is given by $n = \dfrac{(1{\cdot}96)^2 \sigma^2}{E^2}$, where μ and σ^2 are the mean and variance respectively of the population and $\bar{X}$ is the mean of the random sample.*

Solution. By Central Limit Theorem, we know that $\bar{X} \sim N(\mu, \sigma^2/n)$ asymptotically *i.e.,* for large n.

$$\therefore \qquad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ asymptotically } i.e., \text{ for large } n.$$

From normal probability tables, we have

$$P(|Z| \leq 1{\cdot}96) = 0{\cdot}95$$

$$\Rightarrow \quad P\left[\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq 1{\cdot}96\right] = 0{\cdot}95$$

$$\Rightarrow \quad P\left[|\bar{X} - \mu| \leq 1{\cdot}96 \frac{\sigma}{\sqrt{n}}\right] = 0{\cdot}95 \qquad \ldots(*)$$

We are given that

$$P[|\bar{X} - \mu| < E] > 0{\cdot}95 \qquad \ldots(**)$$

From (*) and (**), we have

$$E > \frac{1{\cdot}96\sigma}{\sqrt{n}} \Rightarrow n > \frac{(1{\cdot}96)^2 \sigma^2}{E^2} = \frac{3{\cdot}84\sigma^2}{E^2}$$

Hence minimum sample size n for estimating μ with 95% confidence coefficient is given by $n = 3{\cdot}84\, \sigma^2/E^2$, where E is the permissible error.

Remark. The minimum sample size for estimating μ with confidence coefficient $(1 - \alpha)$ is given by $\sigma^2 z_\alpha^2/E^2$, where z_α is the significant value of Z at level of significance α and E is the permissible error in the estimate.

Arguing similarly, the minimum sample size for estimating population proportion P with confidence coefficient $(1-\alpha)$ is given by $n = PQ\, z_\alpha^2/E^2$, where z_α is the significant value of Z at 'α' level of significance and E is the permissible error in the estimate. If P is unknown, we may use $\hat{P} = p$.

Example 12·18. *The mean muscular endurance score of a random sample of 60 subjects was found to be 145 with a s.d. of 40. Construct a 95% confidence interval for the true mean. Assume the sample size to be large enough for normal approximation. What size of sample is required to estimate the mean within 5 of the true mean with a 95% confidence ?*

[*Calicut Univ. B.Sc. (Main Stat.) 1989*]

Solution. We are given : $n = 60, \bar{x} = 145$ and $s = 40$.

95% confidence limits for true mean (μ) are :

$$\bar{x} \pm 1{\cdot}96\, s/\sqrt{n} \qquad (\sigma^2 = s^2, \text{ since sample is large})$$

$$= 145 \pm \frac{1{\cdot}96 \times 40}{\sqrt{60}} = 145 \pm \frac{78{\cdot}4}{7{\cdot}75} = 145 \pm 10{\cdot}12 = 134{\cdot}88,\ 155{\cdot}12$$

Hence 95% confidence interval for μ is (134·88, 155·12). In the notations of Example 12·17, we have

$$n = \left(\frac{z_\alpha \cdot \sigma}{E}\right)^2 = \left(\frac{1{\cdot}96 \times 40}{5}\right)^2$$

$$[\because z_{0\cdot05} = 1{\cdot}96,\ \hat{\sigma} = s = 40 \text{ and } |\bar{x} - \mu| < 5 = E]$$

$$= (15{\cdot}68)^2 = 245{\cdot}86 \simeq 246.$$

Example 12·19. *The standard deviation of a population is 2·70 inches. Find the probability that in a random sample of size 66 (i) the sample mean will differ from the population mean by 0·75 inch or more and (ii) the sample mean will exceed the population mean by 0·75 inch or more (given that the value of the standard normal probability integral from 0 to 2·25 is 0·4877).*

Solution. Here we are given $n = 66$, $\sigma = 2{\cdot}70$ inches. Since n is large, the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$.

$$\therefore \qquad Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad \ldots(*)$$

We want

(i)
$$P[\,|\bar{x} - \mu| \geq 0{\cdot}75] = 1 - P[\,|\bar{x} - \mu| < 0{\cdot}75]$$

$$= 1 - P\left[\left|\frac{\sigma}{\sqrt{n}} Z\right| < 0{\cdot}75\right] \qquad \text{[From (*)]}$$

$$= 1 - P\left[\,|Z| < 0{\cdot}75\,\frac{\sqrt{n}}{\sigma}\right]$$

$$= 1 - 2\,P\left[\,0 < Z < 0{\cdot}75\,\frac{\sqrt{n}}{\sigma}\right]$$

$$= 1 - 2P\left[0 < Z < 0 \cdot 75 \times \frac{\sqrt{66}}{2 \cdot 70}\right]$$

$$= 1 - 2P\left[0 < Z < \frac{0 \cdot 75 \times 8 \cdot 124}{2 \cdot 70}\right]$$

$$= 1 - 2P[0 < Z < 2 \cdot 25] = 1 - 2 \times 0 \cdot 4877 = 0 \cdot 0246$$

(ii) $P[\bar{x} - \mu > 0 \cdot 75] = P(Z > 0 \cdot 75 \sqrt{n}/\sigma) = P(Z > 2 \cdot 25)$
$= 0 \cdot 5 - P(0 < Z < 2 \cdot 25) = 0 \cdot 5 - 0 \cdot 4877 = 0 \cdot 0123$

Example 12·20. *A normal population has a mean of 0·1 and standard deviation of 2·1. Find the probability that mean of a sample of size 900 will be negative.* **[Delhi Univ. B.Sc. (Stat. Hons.), 1986]**

Solution. Here we are given that $X \sim N(\mu, \sigma^2)$, where $\mu = 0 \cdot 1$ and $\sigma = 2 \cdot 1$ and $n = 900$.

Since $X \sim N(\mu, \sigma^2)$, the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$. The standard normal variate corresponding to $\bar{x}$ is given by :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0 \cdot 1}{2 \cdot 1/30} = \frac{\bar{x} - 0 \cdot 1}{0 \cdot 07}$$

$$\therefore \qquad \bar{x} = 0 \cdot 1 + 0 \cdot 07Z, \text{ where } Z \sim N(0, 1)$$

The required probability p, that the sample mean is negative is given by :

$$p = P(\bar{x} < 0) = P(0 \cdot 1 + 0 \cdot 07\, Z < 0)$$

$$= P\left(Z < \frac{-0 \cdot 10}{0 \cdot 07}\right) = P\,(Z < -1 \cdot 43) = P(Z \geq 1 \cdot 43)$$

$$= 0 \cdot 5 - P(0 < Z < 1 \cdot 43) = 0 \cdot 5 - 0 \cdot 4236 = 0 \cdot 0764$$

(From Normal Probability Tables)·

Example 12·21. *The guaranteed average life of a certain type of electric light bulbs is 1000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90 per cent of the bulbs do not fall short of the guaranteed average by more than 2·5 per cent. What must be the minimum size of the sample ?* **[Madras Univ. B.Sc., Oct. 1991]**

Solution. Here $\mu = 1000$ hours, $\sigma = 125$ hours.

Since we do not want the sample mean to be less than the guaranteed average mean ($\mu = 1000$) by more than 2·5%, we should have

$$\bar{x} > 1000 - 2 \cdot 5\% \text{ of } 1000 \Rightarrow \bar{x} > 1000 - 25 = 975$$

Let n be the given sample size. Then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ since sample is large.}$$

We want $$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{975 - 1000}{125/\sqrt{n}} > -\frac{\sqrt{n}}{5} \qquad (\because \bar{x} > 975)$$

According to the given condition, we have

$$P(Z > -\sqrt{n/5}) = 0{\cdot}90 \Rightarrow P(0 < Z < \sqrt{n/5}) = 0{\cdot}40$$

$\therefore \quad \sqrt{n/5} = 1{\cdot}28$ (From Normal Probability Tables)

$\Rightarrow \quad n = 25 \times (1{\cdot}28)^2 = 41$ (approx)

Example 12·22. *A survey is proposed to be conducted to know the annual earnings of the old Statistics graduates of Delhi University. How large should the sample be taken in order to estimate the mean annual earnings within plus and minus Rs. 1,000 at 95% confidence level ? The standard deviation of the annual earnings of the entire population is known to be Rs. 3,000.*

Solution. We are given : σ = Rs. 3,000.

We want : $P[\,|\bar{x} - \mu| < 1{,}000] = 0{\cdot}95$ ...(*)

We know that, in sampling from normal population or for large samples from any population $\bar{X} \sim N(\mu, \sigma^2/n)$. Hence from normal probability tables, we have :

$$P[\,|Z| \le 1{\cdot}96] = 0{\cdot}95$$

$$\Rightarrow \quad P\left[\,\left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \le 1{\cdot}96\right] = 0{\cdot}95$$

$$\Rightarrow \quad P\left[\,|\bar{x} - \mu| \le 1{\cdot}96 \times (\sigma/\sqrt{n})\right] = 0{\cdot}95 \qquad \text{...(**)}$$

From (*) and (**), we get

$$\frac{1{\cdot}96 \times \sigma}{\sqrt{n}} = 1000 \Rightarrow \frac{1{\cdot}96 \times 3000}{\sqrt{n}} = 1000$$

$$\therefore \quad n = (1{\cdot}96 \times 3)^2 = (5{\cdot}88)^2 = 34{\cdot}56 \simeq 35$$

Aliter. Using Remark to Example 12·17,

$$n = \left(\frac{z_\alpha \cdot \sigma}{E}\right)^2 = \left(\frac{1{\cdot}96 \times 3{,}000}{1{,}000}\right)^2 \simeq 35.$$

12·14. Test of Significance for Difference of Means. Let $\bar{x}_1$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\bar{x}_2$ be the mean of an independent random sample of size n_2 from another population with mean μ_2 and variance σ_2^2. Then, since sample sizes are large,

$$\bar{x}_1 \sim N(\mu_1, \sigma_1^2/n_1) \text{ and } \bar{x}_2 \sim N(\mu_2, \sigma_2^2/n_2)$$

Also $\bar{x}_1 - \bar{x}_2$, being the difference of two independent normal variates is also a normal variate. The Z (*S.N.V.*) corresponding to $\bar{x}_1 - \bar{x}_2$ is given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.\,(\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Under the null hypothesis $H_0 : \mu_1 = \mu_2$, *i.e.*, there is no significant difference between the sample means, we get

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2 = 0;$$

$$V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

the covariance term vanishes, since the sample means $\bar{x}_1$ and $\bar{x}_2$ are independent.

Thus under $H_0 : \mu_1 = \mu_2$, the test statistic becomes (for large samples),

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \sim N(0, 1) \qquad \ldots(12 \cdot 11)$$

Remarks 1. If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.*, if the samples have been drawn from the populations with common S.D. σ, then under $H_0 : \mu_1 = \mu_2$,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{(1/n_1 + 1/n_2)}} \sim N(0, 1) \qquad \ldots[12 \cdot 11(a)]$$

2. If in (12·11*a*), σ is not known, then its estimate based on the sample variances is used. If the sample sizes are not sufficiently large, then an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)},$$

since
$$E(\hat{\sigma}^2) = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)E(S_1^2) + (n_2 - 1)E(S_2^2)\right]$$
$$= \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2\right] = \sigma^2$$

But since sample sizes are large, $S_1^2 \simeq s_1^2$, $S_2^2 \simeq s_2^2$, $n_1 - 1 \simeq n_1$, $n_2 - 1 \simeq n_2$. Therefore in practice, for large samples, the following estimate of σ^2 without any serious error is used :

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \qquad \ldots[12 \cdot 11(b)]$$

However, if sample sizes are small, then a small sample test, *t*-test for difference of means (*c.f.* Chapter 14) is to be used.

3. If $\sigma_1^2 \neq \sigma_2^2$ and σ_1 and σ_2 are not known, then they are estimated from sample values. This results in some error, which is practically immaterial, if samples are large. These estimates for large samples are given by

$$\left.\begin{aligned} \hat{\sigma}_1^2 &= S_1^2 \simeq s_1^2 \\ \hat{\sigma}_2^2 &= S_2^2 \simeq s_2^2 \end{aligned}\right\} \quad \text{(since samples are large).}$$

In this case, (12·11) gives

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \sim N(0, 1) \qquad \ldots[12 \cdot 11(c)]$$

Example 12·23. *The means of two single large samples of 1000 and 2000 members are 67·5 inches and 68·0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2·5 inches ? (Test at 5% level of significance).*

Solution. We are given :

$$n_1 = 1000, n_2 = 2000;\ \bar{x}_1 = 67{\cdot}5 \text{ inches},\ \bar{x}_2 = 68{\cdot}0 \text{ inches}.$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$ and $\sigma = 2{\cdot}5$ inches, *i.e.*, the samples have been drawn from the same population of standard deviation 2·5 inches.

Alternative Hypothesis, $H_1 : \mu_1 \neq \mu_2$ (Two tailed.)

Test Statistic. Under H_0, the test statistic is (since samples are large)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

Now $$Z = \frac{67{\cdot}5 - 68{\cdot}0}{2{\cdot}5 \times \sqrt{\frac{1}{1000} + \frac{1}{2000}}} = \frac{-0{\cdot}5}{2{\cdot}5 \times 0{\cdot}0387} = -5{\cdot}1$$

Conclusion. Since $|Z| > 3$, the value is highly significant and we reject the null hypothesis and conclude that samples are certainly not from the same population with standard deviation 2·5.

Example 12·24. *In a survey of buying habits, 400 women shoppers are chosen at random in super market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.*

Solution. In the usual notations, we are given that

$$n_1 = 400, \qquad \bar{x}_1 = \text{Rs. } 250, \qquad s_1 = \text{Rs. } 40$$

$$n_2 = 400, \qquad \bar{x}_2 = \text{Rs. } 220 \qquad s_2 = \text{Rs. } 55$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.*, the average weekly food expenditures of the two populations of shoppers are equal.

Alternative Hypothesis, $H_1 : \mu_1 \neq \mu_2$. (Two-tailed)

Test Statistic. Since samples are large, under H_0, the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} \sim N(0, 1)$$

Since σ_1 and σ_2, the population standard deviations are not-known, we can take for large samples (*c.f.* § 12·15, Remark 3) :

$$\hat{\sigma}_1^2 = s_1^2 \quad \text{and} \quad \hat{\sigma}_2^2 = s_2^2$$

and then Z is given by

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{250 - 220}{\sqrt{\left\{\frac{(40)^2}{400} + \frac{(55)^2}{400}\right\}}} = 8{\cdot}82 \text{ (approx.)}$$

Conclusion. Since $|Z|$ is much greater than 2·58, the null hypothesis ($\mu_1 = \mu_2$) is rejected at 1% level of significance and we conclude that the average weekly expenditures of two populations of shoppers in markets A and B differ significantly.

Example 12·25. *The average hourly wage of a sample of 150 workers in a plant 'A' was Rs. 2·56 with a standard deviation of Rs. 1·08. The average wage of a sample of 200 workers in plant 'B' was Rs. 2·87 with a standard deviation of Rs. 1·28. Can an applicant safely assume that the hourly wages paid by plant 'B' are higher than those paid by plant 'A' ?*

Solution. Let X_1 and X_2 denote the hourly wages (in *Rs.*) of workers in plant A and plant B respectively. Then we are given :

$$n_1 = 150,\ \bar{x}_1 = 2\cdot56,\ s_1 = 1\cdot08 = \hat{\sigma}_1$$

$$n_2 = 200,\ \bar{x}_2 = 2\cdot87,\ s_2 = 1\cdot28 = \hat{\sigma}_2$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.,* there is no significant difference between the mean level of wages of workers in plant A and plant B.

Alternative hypothesis, $H_1 : \mu_2 > \mu_1$ *i.e.,* $\mu_1 < \mu_2$ (Left-tailed test)

Test Statistic. Under H_0, the test statistic (for large samples) is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$\therefore \quad Z = \frac{2\cdot56 - 2\cdot87}{\sqrt{\left\{\frac{(1\cdot08)^2}{150} + \frac{(1\cdot28)^2}{200}\right\}}} = \frac{-0\cdot31}{\sqrt{0\cdot016}} = \frac{-0\cdot31}{0\cdot126} = -2\cdot46.$$

Critical region. For a one-tailed test, the critical value of Z at 5% level of significance is 1·645. The critical region for left-tailed test thus consists of all values of $Z \leq -1\cdot645$.

Conclusion. Since calculated value of Z (−2·46) is less than critical value (−1·645), it is significant at 5% level of significance. Hence the null hypothesis is rejected at 5% level of significance and we conclude that the average hourly wages paid by plant 'B' are certainly higher than those paid by plant 'A'.

Example 12·26. *In a certain factory there are two independent processes manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 ozs. with a standard deviation of 12 ozs. while the corresponding figures in a sample of 400 items from the other process are 124 and 14. Obtain the standard error of difference between the two sample means. Is this difference significant ? Also find the 99% confidence limits for the difference in the average weights of items produced by the two processes respectively.*

Solution. We have

$$\left.\begin{aligned} n_1 &= 250,\ \bar{x}_1 = 120 \text{ oz.},\ s_1 = 12 \text{ oz.} = \hat{\sigma}_1 \\ n_2 &= 400,\ \bar{x}_2 = 124 \text{ oz.},\ s_2 = 14 \text{ oz.} = \hat{\sigma}_2 \end{aligned}\right\}, \text{ (since samples are large).}$$

$$S.E.\ (\bar{x}_1 - \bar{x}_2) = \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)} = \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$$

$$= \sqrt{\left(\frac{144}{250} + \frac{196}{400}\right)} = \sqrt{(0.576 + 0.490)} = 1.034$$

Null Hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.*, the sample means do not differ significantly.

Alternative Hypothesis, $H_1 : \mu_1 \neq \mu_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{S.E.\ (\bar{x}_1 - \bar{x}_2)} = \frac{120 - 124}{1.034} \sim N(0, 1)$$

$$\therefore \qquad |Z| = \frac{4}{1.034} = 3.87$$

Conclusion. Since $|Z| > 3$, the null hypothesis is rejected and we conclude that there is significant difference between the sample means.

99% confidence limits for $|\mu_1 - \mu_2|$, *i.e.*, for the difference in the average weights of items produced by two processes, are

$$|\bar{x}_1 - \bar{x}_2| \pm 2.58\ S.E.\ (\bar{x}_1 - \bar{x}_2) = 4 \pm 2.58 \times 1.034$$

$$= 4 \pm 2.67 \text{ (approx.)} = 6.67 \text{ and } 1.33$$

$$\therefore \qquad 1.33 < |\mu_1 - \mu_2| < 6.67$$

Example 12·27. *The mean height of 50 male students who showed above average participation in college athletics was 68·2 inches with a standard deviation of 2·5 inches; while 50 male students who showed no interest in such participation had a mean height of 67·5 inches with a standard deviation of 2·8 inches.*

(i) Test the hypothesis that male students who participate in college athletics are taller than other male students.

(ii) By how much should the sample size of each of the two groups be increased in order that the observed difference of 0·7 inches in the mean heights be significant at the 5% level of significance.

Solution. Let X_1 and X_2 denote the height (in inches) of athletic participants and non-athletic participants respectively. In the usual notations, we are given :

$$n_1 = 50, \quad \bar{x}_1 = 68.2, \quad s_1 = 2.5 \ ; n_2 = 50, \quad \bar{x}_2 = 67.5, \quad s_2 = 2.8$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$.

Alternative hypothesis, $H_1 : \mu_1 > \mu_2$ (Right-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$\therefore \qquad Z = \frac{68.2 - 67.5}{\sqrt{\left\{\frac{(2.5)^2}{50} + \frac{(2.8)^2}{50}\right\}}} = \frac{0.7}{\sqrt{0.282}} = \frac{0.7}{0.53} = 1.32$$

For a right-tailed test, the critical (significant) value of Z at 5% level of significance is 1·645.

(*i*) Since the calculated value of $Z(1\cdot32)$ is less than the critical value (1·645), it is not significant at 5% level of significance. Hence the null hypothesis is accepted and we conclude that the college athletes are not taller than other male students.

(*ii*) The difference between the mean heights of two groups, each of size n will be significant at 5% level of significance if $Z \geq 1\cdot645$

$$\Rightarrow \qquad \frac{68\cdot2 - 67\cdot5}{\sqrt{\left\{\dfrac{(2\cdot5)^2}{n} + \dfrac{(2\cdot8)^2}{n}\right\}}} \geq 1\cdot645$$

$$\Rightarrow \qquad \frac{0\cdot7}{\sqrt{14\cdot09/n}} \geq 1\cdot645 \Rightarrow \frac{0\cdot7}{3\cdot754/\sqrt{n}} \geq 1\cdot645$$

$$\Rightarrow \qquad n \geq \left(\frac{1\cdot645 \times 3\cdot754}{0\cdot7}\right)^2 = (8\cdot8219)^2 = 77\cdot83 \simeq 78$$

Hence the sample size of each of the two groups should be increased by at least $78 - 50 = 28$, in order that the difference between the mean heights of the two groups is significant.

12·15. Test of Significance for the Difference of Standard Deviations. If s_1 and s_2 are the standard deviations of two independent samples, then under null hypothesis, $H_0 : \sigma_1 = \sigma_2$, *i.e.*, *i.e.*, the sample standard deviations don't differ significantly, the statistic

$$Z = \frac{s_1 - s_2}{\text{S.E.}(s_1 - s_2)} \sim N(0, 1) \text{ for large samples.}$$

But in case of large samples, the S.E of the difference of the sample standard deviations is given by

$$\text{S.E.}(s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

$$\therefore \qquad Z = \frac{s_1 - s_2}{\sqrt{\left(\dfrac{\sigma_1^2}{2n_1} + \dfrac{\sigma_2^2}{2n_2}\right)}} \sim N(0, 1) \qquad \ldots(12\cdot12)$$

σ_1^2 and σ_2^2 are usually unknown and for large samples, we use their estimates given by the corresponding sample variances. Hence the test statistic reduces to

$$Z = \frac{s_1 - s_2}{\sqrt{\left(\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}\right)}} \sim N(0, 1) \qquad \ldots(12\cdot13)$$

Example 12·18. *Random samples drawn from two countries gave the following data relating to the heights of adult males :*

	Country A	*Country B*
Mean height (in inches)	*67·42*	*67·25*
Standard deviation (in inches)	*2·58*	*2·50*
Number in samples	*1000*	*1200*

(i) Is the difference between the means significant ?

(ii) Is the difference between the standard deviations significant ?

Solution. We are given :

$$n_1 = 1000, \quad \bar{x}_1 = 67{\cdot}42 \text{ inches}, \quad s_1 = 2{\cdot}58 \text{ inches},$$

$$n_2 = 1200, \quad \bar{x}_2 = 67{\cdot}25 \text{ inches}, \quad s_2 = 2{\cdot}50 \text{ inches}.$$

As in the last examples (since sample sizes are large), we can take

$$\hat{\sigma}_1 = s_1 = 2{\cdot}58, \quad \hat{\sigma}_2 = s_2 = 2{\cdot}50$$

(*i*) $H_0 : \mu_1 = \mu_2$, *i.e.*, the sample means do not differ significantly.

$$H_1 : \mu_1 \neq \mu_2 \text{ (Two tailed).}$$

Under the *Null hypothesis* H_0, the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \sim N(0, 1), \text{ since samples are large.}$$

Now $$Z = \frac{67{\cdot}42 - 67{\cdot}25}{\sqrt{\left\{\frac{(2{\cdot}58)^2}{1000} + \frac{(2{\cdot}50)^2}{1200}\right\}}} = \frac{0{\cdot}17}{\sqrt{\left(\frac{6{\cdot}66}{1000} + \frac{6{\cdot}25}{1200}\right)}} = 1{\cdot}56$$

Conclusion. Since $|Z| < 1{\cdot}96$, null hypothesis may be accepted at 5% level of significance and we may conclude that there is no significant difference between the sample means.

(*ii*) Under H_0 : that there is no significant difference between sample standard deviations,

$$Z = \frac{s_1 - s_2}{\text{S.E. }(s_1 - s_2)} \sim N(0, 1), \text{ since samples are large.}$$

$$\text{Now S.E. }(s_1 - s_2) = \sqrt{\left(\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}\right)} = \sqrt{\left(\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}\right)},$$

if σ_1 and σ_2 are not known and $\hat{\sigma}_1 = s_1$, $\hat{\sigma}_2 = s_2$.

$$\therefore \quad S.E.\ (s_1 - s_2) = \sqrt{\left\{\frac{(2{\cdot}58)^2}{2 \times 1000} + \frac{(2{\cdot}50)^2}{2 \times 1200}\right\}} = 0{\cdot}07746$$

Hence $$Z = \frac{2{\cdot}58 - 2{\cdot}50}{0{\cdot}07746} = \frac{0{\cdot}08}{0{\cdot}07746} = 1{\cdot}03$$

Conclusion. Since $|Z| < 1{\cdot}96$, the data don't provide us any evidence against the null hypothesis which may be accepted at 5% level of significance. Hence the sample standard deviations do not differ significantly.

Example 12·29. *Two populations have their means equal, but S.D. of one is twice the other. Show that in the samples of size 2000 from each drawn under simple sampling conditions, the difference of means will, in all probability, not exceed $0{\cdot}15\sigma$, where σ is the smaller S.D. What is the probability that the difference will exceed half this amount ?*

Solution. Let the standard deviations of the two populations be σ and 2σ respectively and let μ be the mean of each of the two populations. Also we are given $n_1 = n_2 = 2000$. If $\bar{x}_1$ and $\bar{x}_2$ be the two sample means then, since samples are large,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.\ (\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Now $\quad E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu - \mu = 0$ and

$$S.E.\ (\bar{x}_1 - \bar{x}_2) = \sqrt{\left\{\frac{\sigma^2}{n_1} + \frac{(2\sigma)^2}{n_2}\right\}} = \sigma.\sqrt{\left(\frac{1}{2000} + \frac{4}{2000}\right)} = 0{\cdot}05\sigma$$

$$\therefore \quad Z = \frac{\bar{x}_1 - \bar{x}_2}{S.E.\ (\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Under simple sampling conditions, we should in all probability have

$$|Z| < 3 \Rightarrow |\bar{x}_1 - \bar{x}_2| < 3\ S.E.\ (\bar{x}_1 - \bar{x}_2)$$

$$\Rightarrow \quad |\bar{x}_1 - \bar{x}_2| < 0{\cdot}15\sigma,$$

which is the required result.

We want $\quad p = P\left[\,|\bar{x}_1 - \bar{x}_2| > \frac{1}{2} \times 0{\cdot}15\sigma\right]$

$$\therefore \quad p = P\left[0{\cdot}05\sigma\,|Z| > 0{\cdot}075\sigma\right] \quad \left[\because\ Z = \frac{\bar{x}_1 - \bar{x}_2}{0{\cdot}05\sigma} \sim N\ (0, 1)\right]$$

$$= P\ [\,|Z| > 1{\cdot}5\,] = 1 - P\ [\,|Z| \le 1{\cdot}5\,]$$

$$= 1 - 2P\ (0 \le Z \le 1{\cdot}5) = 1 - 2 \times 0{\cdot}4332 = 0{\cdot}1336$$

EXERCISE 12·2

1. Define sampling distribution and standard error. Obtain standard error of mean when population is large.

2. Find the standard error of a linear function of a number of variables. Deduce the standard error of the mean of n uncorrelated variables following the same distribution.

3. Derive the expressions for the standard error of

(*i*) the mean of a random sample of size n, and

(*ii*) the difference of the means of two independent random samples of sizes n_1 and n_2.

4. (*a*) What is meant by a statistical hypothesis ? What are the two types of errors of decision that arise in testing a hypothesis ? Briefly explain how a statistical hypothesis is tested.

The manufacturer of television tubes knows from past experience that the average life of a tube is 2,000 hours with a standard deviation of 200 hours. A sample of 100 tubes has an average life of 1950 hours. Test at the 0·05 level of significance if this sample came from a normal population of mean 2,000 hours.

State your null and alternative hypothesis and indicate clearly whether a one-tail or a two-tail test is used and why ? Is the result of the test significant ?

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1990*]

(*b*) A sample of 100 items, drawn from a universe with mean value 64 and S.D. 3 has a mean value 63·5. Is the difference in the means significant ? What will be your inference, if the sample had 200 items ?

[*Madras Univ. B.E., Nov. 1990*]

(*c*) A sample of 400 individuals is found to have a mean height of 67·47 inches. Can it be reasonably regarded as a sample from a large population with mean height of 67·39 inches and standard deviation 1·30 inches ?

Ans. Yes, $Z = 1 \cdot 23$.

(*d*) The mean breaking strength of cables supplied by a manufacturer is 1800 with a standard deviation 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cables has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 0·01 level of significance ?

Ans. $H_0 : \mu = 1800, H_1 : \mu > 1800, Z = 3 \cdot 535$.

(*e*) An ambulance service claims that it takes on the average 8·9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licenses ambulance services has them timed on 50 emergency calls, getting a mean of 9·3 minutes with a standard deviation of 1·6 minutes. What can they conclude at the level of significance $\alpha = 0 \cdot 05$?

Ans. $Z = 1 \cdot 768$.

(*f*) A paper mill in U.P. has agreed to buy waste paper for recycling from a waste collection firm, under the agreement that the waste collection firm will supply the waste paper in packages of 300 kg each, for which the paper mill will pay by the package. To speed up their work the waste collection firm is making packages by some *approximation* procedure. The paper mill does not object to this procedure as long as it gets 300 kg. per package on the average. The waste collection firm has an interest not to exceed 300 kg. per package, because it is not being paid for more, and not to go under 300 kg. because the paper mill might terminate the agreement if it does. To estimate the mean weight of waste paper per package, the waste collection firm weighed 75 randomly selected packages and found that the mean weight was 290 kg and standard deviation was 15 kg. Can we infer that the mean weight per package in the entire supply was 300 kg ? [*Delhi Univ. M.A. (Eco.), 1987*]

Ans. $H_0 : \mu = 300$ kg; $H_1 : \mu \neq 300$ kg. (Two-tailed).

$$Z = \frac{290 - 300}{15/\sqrt{75}} = 5 \cdot 77 \text{ ; Significant.}$$

(*g*) The wages of a factory's workers are assumed to be normally distributed with mean μ and variance 25. A random sample of 25 workers gives the total wages equal to 1250 units.

Test the hypothesis : $\mu = 52$, against the alternative : $\mu = 49$, at 1% level of significance.

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2\cdot32} \exp\left(-\tfrac{1}{2}u^2\right) du \doteq 0\cdot01.$$

[*Calcutta Univ. B.Sc.(Maths. Hons.), 1988*]

Ans. $H_0 : \mu = 52,\ H_1 : \mu = 49 < 52$, (Left-tailed test).

$Z = -2$, Not significant.

5. (*a*) A sample of 450 items is taken from a population whose standard deviation is 20. The mean of the sample is 30. Test whether the sample has come from a population with mean 29. Also calculate the 95% confidence limits for the population mean.

(*b*) A sample of 400 observations has mean 95 and standard deviation 12. Could it be a random sample from a population with mean 98 ? What can be the maximum value of the population mean ?

6. (*a*) If the mean age at death of 64 men engaged in an occupation is 52·4 years with standard deviation of 10·2 years, what are the 98% confidence limits for the mean age of all men in that population ?

[*Calicut Univ. B.Sc. (Subs.), 1989*]

(*b*) The weights of 1500 ball bearings are normally distributed with mean 22·40 and standard deviation 0·048. If 300 random samples of size 36 each are drawn from this population, determine the expected mean and standard deviation of the sampling distribution of means, if sampling is done with replacement.

How many of the random samples in the above problem would have their means between 22·39 and 22·41 ? **[*Madras Univ. B.E., April 1989*]**

Hint. $E(\overline{X}) = \mu = 22\cdot40;\ S.E.(\overline{X}) = \sigma/\sqrt{n} = 0\cdot048/\sqrt{36} = 0\cdot008$

Required number of samples (out of 300) is : $300 \times P(22\cdot39 < \overline{X} < 22\cdot41)$

$$= 300 \times P\left(\frac{22\cdot39 - 22\cdot40}{0\cdot008} < Z < \frac{22\cdot41 - 22\cdot40}{0\cdot008}\right);\ Z \sim N(0,\ 1)$$

$$= 300 \times P(-1\cdot25 < Z < 1\cdot25) = 600 \times P(0 < Z < 1\cdot25) \simeq 237$$

7. (*a*) A random sample of 500 is drawn from a large number of freshly minted coins. The mean weight of the coins in the sample is 28·57 gm. and the standard deviation is 1·25 gm. What are the limits which have a 49 to 1 chance of including the mean weight of all the coins ? How large a sample would have to be drawn to make these limits differ by only 0·1 gm, assuming that the standard deviation of the whole distribution is 1·25 gm.

(*b*) A research worker wishes to estimate the mean of a population by using sufficiently large sample. The probability is 0·95 that the sample mean will not differ from the true mean of a normal population by more than 25% of the standard deviation. How large a sample should be taken ? (**Ans.** $n = 62$.)

8. (*a*) A normal distribution has mean 0·5 and standard deviation 2·5. Find :

(*i*) The probability that the mean of a random sample of size 16 from the population is positive.

(*ii*) The probability that the mean of a sample of size 90 from the population will be negative.

(*b*) The mean of a certain normal distribution is equal to the standard error of the mean of a random sample of 100 from that distribution. Find the probability, (in terms of an integral), that the mean of a sample of 25 from the distribution will be negative. (**Ans.** 0·3085.)

(*c*) The average value $\bar{x}$ of a random sample of observations from a certain population is normally distributed with mean 20 and standard deviation $5/\sqrt{n}$. How large a sample should be drawn in order to have a probability of at least 0·90 that $\bar{x}$ will lie between 18 and 22. (***Karnataka Univ. B.E. 1991***)

9. (*a*) From a population of 169 units it is desired to choose a simple random sample of size *n*. If the population standard deviation is 2, determine the smallest '*n*' for which the probability that the sample mean differs from the population mean by more than 0·75 is controlled at 0·05.

(*b*) An economist would like to estimate the mean income (μ) in a large city. He has decided to use the sample mean as an estimate of μ and would like to ensure that the error in estimation is not more than Rs. 100 with probability 0·90. How large a sample should he take if the standard deviation is known to be Rs. 1,000 ? [***Delhi Univ. M.A. (Eco.), 1986***]

$$\textbf{Ans.}\quad n=\left[\frac{z_\alpha \cdot \sigma}{E}\right]^2=\left[\frac{1\cdot645\times1000}{100}\right]^2=270\cdot6\simeq271$$

(*c*) The management of a manufacturing firm wishes to determine the average time required to complete a certain manual operation. There should be 0·95 confidence that the error in the estimate will not exceed 2 minutes.

What sample size is required if the standard deviation of the time needed to complete the manual operation is estimated by a time and motion study expert as (*i*) 10 minutes, (*ii*) 16 minutes ? Explain intuitively (without referring to the formula) why the sample size is large in (*ii*) than in (*i*).

(*Given* $Z_{0\cdot975}=1\cdot96$ and $Z_{0\cdot95}=1\cdot645$)

[***Delhi Univ. M.C.A., 1987***]

$$\textbf{Ans.}\quad (i)\ n_1=\left(\frac{z_\alpha \cdot \sigma}{E}\right)^2=\left(\frac{1\cdot96\times10}{2}\right)^2=96,\quad (ii)\ n_2=\left(\frac{1\cdot96\times16}{2}\right)^2=246.$$

10. (*a*) Two populations have the same mean, but the standard deviation of one is twice that of the other. Show that in samples of 500 each drawn under simple random conditions, the difference of the means will, in all probability, not exceed $0\cdot3\sigma$, where σ is the smaller standard deviation, and assuming the distribution of the difference of the means to be normal, find the probability that it exceeds half that amount. (**Ans.** 0·1336.)

(*b*) A simple sample of heights of 6,400 Englishmen has a mean of 67·85 inches and S.D. 2·56 inches, while a simple sample of heights of 1,600 Australians has a mean of 68·55 inches and a S.D. of 2·52 inches. Do the data indicate that Australians are, on the average, taller than Englishmen ?

Ans. $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 < \mu_2, Z = 9{\cdot}2$, (significant).

(*c*) In a random sample of 500, the mean is found to be 20. In another independent sample of 400, the mean is 15. Could the samples have been drawn from the same population with standard deviation 4 ?

11. (*a*) The following table presents data on the values of a harvested crop stored in the open and inside a godown :

	Sample size	*Mean*	$\Sigma(x-\bar{x})^2$
Outside	40	117	8,685
Inside	100	132	27,315

Assuming that the two samples are random and they have been drawn from normal populations with equal variances, examine if the mean value of the harvested crop is affected by weather conditions.

Ans. $Z = 0{\cdot}342$; Not significant.

(*b*) Samples of students were drawn from two universities and from their weights in kgm., means and standard deviations are calculated. Make a large sample test to test the significance of the difference between the means.

	Mean	*S.D.*	*Size of sample*
University *A*	55	10	400
University *B*	57	15	100

Ans. $Z = 1{\cdot}2648$; Not significant.

(*c*) A storekeeper wanted to buy a large quantity of light bulbs from two brands labelled 'one' and 'two'. He bought 100 bulbs from each brand and found by testing that brand 'one' had mean lifetime of 1120 hours and the standard deviation of 75 hours; and brand 'two' had mean lifetime of 1062 hours and standard deviation of 82 hours. Examine whether the difference of means is significant.

12. The mean yield of two sets of plots and their variability are as given below. Examine

(*i*) whether the difference in the mean yields of the two sets of plots is significant, and

(*ii*) whether the difference in the variability in yields is significant.

	Set of 40 plots	*Set of 60 plots*
Mean yield per plot	1258 lb.	1243 lb.
S.D. per plot	34 lb.	28 lb.

Ans. (*i*) $Z = 2{\cdot}3$, (*ii*) $Z = 1{\cdot}3$.

13. (*a*) In a survey of incomes of two classes of workers, two random samples gave the following details. Examine whether the differences between the (*i*) means and (*ii*) the standard deviations, are significant.

Sample	*Size*	*Mean annual income (in rupees)*	*Standard deviation (in rupees)*
I	100	582	24
II	100	546	28

Examine also whether the first sample could have come from a population with annual mean income of 500 rupees.

(*b*) The electric light tubes of manufacturer *A* have a lifetime of 1400 hours, with a standard deviation of 200 hours, while of manufacture *B* have a mean lifetime of 1200 hours with a standard deviation of 100 hours. If random samples of 125 tubes of each batch are tested, what is the probability that the brand *A* tubes will have a mean time which is at least (*i*) 160 hours more than the brand *B* tubes, and (*ii*) 250 hours more than the brand *B* tubes ?

Hint. Under the assumption of normal population, the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ would have mean; $\mu_1 - \mu_2 = 1400 - 1200 = 200$ hours and standard deviation :

$$\text{S.E.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} = \sqrt{\left\{\frac{(100)^2}{125} + \frac{(200)^2}{125}\right\}} = 20 \text{ hours.}$$

(*i*) The required probability is given by :

$$P\{(\bar{x}_1 - \bar{x}_2) \geq 160\} = P\left[\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{S.E.}(\bar{x}_1 - \bar{x}_2)} \geq \frac{160 - 200}{20}\right]$$

$$= P(Z \geq -2) = 0{\cdot}5 + P(-2 < Z < 0)$$

(*ii*) The required probability is given by :

$$P\{(\bar{x}_1 - \bar{x}_2) \geq 250\} = P(Z \geq 2{\cdot}5) = 0{\cdot}5 - P(0 < Z < 2{\cdot}5)$$

$$= 0{\cdot}5 - 0{\cdot}4938 = 0{\cdot}0062$$

14. A random sample of 1,200 men from one State gives the mean pay as Rs. 400 p.m.with a standard deviation of Rs. 60, and a random sample of 1,000 men from another State gives the mean pay as Rs. 500 p.m., with a standard deviation of Rs. 80.

Discuss, (stating clearly the result or theorem used), whether the mean levels of pay of men from the two States differ significantly.

15. (*a*) A normal population has a mean 0·1 and a standard deviation 2·1. Find the probability that the mean of a sample of size 900 will be negative, it being given that the probability that the absolute value of a standard normal variate exceeds 1·43 is 0·153.

(*b*) A random sample of 100 articles selected from a batch of 2,000 articles shows that the average diameter of the articles is 0·354 with a standard deviation 0·048. Find 95% confidence interval for the average of this batch of 2,000 articles.

Hint. We are given $n = 100, N = 2{,}000, \bar{x} = 0{\cdot}354, s = 0{\cdot}048$.

The Standard Error of sample mean $\bar{x}$ in random sampling from the batch of $N = 2{,}000$ is given by : [c.f. (16·23)].

$$S.E.(\bar{x}) = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{N-n}{N-1}} \times \frac{s}{\sqrt{n}} \quad (\because \hat{\sigma} = s, \text{ for large } n)$$

$$= \sqrt{\frac{2000 - 100}{2000 - 1}} \times \frac{0{\cdot}048}{\sqrt{100}} = 0{\cdot}00468$$

Hence 95% confidence limits for μ are given by :

$$\bar{x} \pm 1.96\ S.E.(\bar{x}) = 0.354 \pm 1.96 \times 0.00468 = (0.3448,\ 0.3632)$$

16. (*a*) Explain the terms :

(*i*) Statistic and Parameter

(*ii*) Sampling distribution of a statistic, and

(*iii*) Standard error of a statistic.

(*b*) Explain why a random sample of size 30 is to be preferred to a random sample of size 25 to estimate the population mean.

17. (*a*) Obtain the expressions for the standard error of sampling distributions of : (*i*) sample mean ($\bar{x}$), and (*ii*) sample variance (s^2), in random sampling from a large population. Assume that n, the sample size, is large.

(*b*) Let $X_1, X_2, \ldots, X_n$ be a random sample from a population which has a finite fourth moment $\mu_r = E(X_i - \mu)^r$, $r = 4$; $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$; and

let : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Show that : (*i*) $S^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2$,

(*ii*) $\text{Var}(S^2) = \frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right]$,

(*iii*) $\text{Cov}(\bar{X}, S^2) = \mu_3/n$

Exact Sampling Distributions (Chi-square Distribution)

13·1. Chi-Square Variate *(Pronounced as Ki - Sky without S)*. The square of a standard normal variate is known as a chi-square variate with 1 degree of freedom (d.f.)

Thus if $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma} \sim N(0, 1)$

and $\quad Z^2 = \left(\dfrac{X-\mu}{\sigma}\right)^2$, is a chi-square variate with 1 d.f.

In general, if X_i, $(i = 1, 2, \ldots, n)$ are n independent normal variates with mean μ_i and variance σ_i^2, $(i = 1, 2, \ldots, n)$, then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2, \text{ is a chi-square variate with } n \text{ d.f.} \qquad \ldots(13\cdot1)$$

13·2. Derivation of the Chi-square Distribution.

First Method—Method of Moment Generating Function.

If X_i, $(i = 1, 2, \ldots, n)$ are independent $N(\mu_i, \sigma_i^2)$, we want the distribution of

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} U_i^2, \quad \text{where } U_i = \frac{X_i - \mu_i}{\sigma_i}$$

Since X_i's are independent, U_i's are also independent.

$$M_{\chi^2}(t) = M_{\Sigma U_i^2}(t) = \prod_{i=1}^{n} M_{U_i^2}(t) = [M_{U_i^2}(t)]^n,$$

since U_i's $\sim N(0, 1)$ are identically distributed.

Now

$$M_{U_i^2}(t) = E[\exp\{tU_i^2\}] = \int_{-\infty}^{\infty} \exp(tu_i^2) f(x_i) dx_i$$

$$= \int_{-\infty}^{\infty} \exp(tu_i^2) \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x_i - \mu)^2/2\sigma^2\} dx_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(tu_i^2) \exp(-u_i^2/2)\, du_i \qquad \left[u_i = \frac{x_i - \mu}{\sigma}\right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{1-2t}{2}\right)u_i^2\right\} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}\ ^*}{\left(\frac{1-2t}{2}\right)^{1/2}} = (1-2t)^{-1/2}$$

$$\therefore \quad M_{\chi^2}(t) = (1-2t)^{-n/2} \qquad \ldots(13\cdot1a)$$

which is the m.g.f. of a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$.

Hence by uniqueness theorem of m.g.f.'s,

$$\chi^2 = \sum_i^n \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

is a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$.

$$\therefore \quad dP(\chi^2) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} \cdot [\exp(-\tfrac{1}{2}\chi^2)]\ (\chi^2)^{(n/2)-1} d\chi^2$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} [\exp(-\chi^2/2)]\ (\chi^2)^{(n/2)-1} d\chi^2,\ 0 \le \chi^2 < \infty \quad \ldots(13\cdot2)$$

which is the required probability density function of chi-square distribution with n degrees of freedom.

Remarks 1. If a random variable X has a chi-square distribution with n d.f., we write $X \sim \chi^2_{(n)}$ and its *p.d.f.* is given by :

$$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-x/2}\ x^{(n/2)-1}\ ;\ 0 \le x < \infty \qquad \ldots(13\cdot2a)$$

2. If $X \sim \chi^2_{(n)}$, then $(X/2) \sim \gamma(n/2)$.

Proof. The p.d.f. of $Y = \frac{1}{2}X$, is given by :

$$g(y) = f(x) \cdot \left|\frac{dx}{dy}\right|$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-y} \cdot (2y)^{(n/2)-1} \cdot 2$$

$$= \frac{1}{\Gamma(n/2)} e^{-y}\ y^{(n/2)-1}\ ;\ 0 \le y < \infty$$

$$\Rightarrow \qquad Y = (X/2) \sim \gamma(n/2)$$

Second Method—Method of Induction

If X_i is a $N(0, 1)$, then $X_i^2/2$ is a $\gamma(1/2)$ so that X_i^2 is a χ^2-variate with d.f. 1.

---

* $\int_{-\infty}^{\infty} \exp(-a^2x^2)\,dx = \frac{\sqrt{\pi}}{a}$

If X_1 and X_2 are independent standard normal variates then $X_1^2 + X_2^2$ is a chi-square variate with 2 d.f. which may be proved as follows :

The joint probability differential of X_1 and X_2 is given by :

$$dP(x_1, x_2) = f(x_1, x_2)\, dx_1\, dx_2 = f_1(x_1) f_2(x_2) dx_1\, dx_2$$

$$= \frac{1}{2\pi} \exp\left\{-(x_1^2 + x_2^2)/2\right\} dx_1\, dx_2, -\infty < (x_1, x_2) \le \infty$$

Let us now transform to polar co-ordinates by the substitution $x_1 = r\cos\theta$, $x_2 = r\sin\theta$. Jacobian of transformation J is given by

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r} \\ \frac{\partial x_1}{\partial \theta} & \frac{\partial x_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{vmatrix} = r$$

Also we have $r^2 = x_1^2 + x_2^2$ and $\tan\theta = x_2/x_1$. As x_1 and x_2 range from $-\infty$ to $+\infty$, r varies from 0 to ∞ and θ from 0 to 2π. The joint probability differential of r and θ now becomes

$$dG(r, \theta) = \frac{1}{2\pi} \exp(-r^2/2)\, r\, dr d\theta\,;\ 0 \le r \le \infty, 0 \le \theta \le 2\pi$$

Integrating over θ, the marginal distribution of r is given by

$$dG_1(r) = \int_0^{2\pi} dG(r, \theta) = r\exp(-r^2/2)\, dr \left[\frac{\theta}{2\pi}\right]_0^{2\pi}$$

$$= \exp(-r^2/2)\, r\, dr$$

$$\Rightarrow \quad dG_1(r^2) = \frac{1}{2}\exp(-r^2/2)\, dr^2$$

$$= \frac{1}{\Gamma(1)}\exp(-r^2/2)\,(r^2/2)^{1-1}\, d(r^2/2)$$

Thus $\dfrac{r^2}{2} = \dfrac{X_1^2 + X_2^2}{2}$ is a $\gamma(1)$ variate and hence $r^2 = X_1^2 + X_2^2$ is a χ^2-variate with 2 d.f.

For n variables X_i, $(i = 1, 2, \ldots, n)$ we transform $(X_1, X_2, \ldots, X_n)$ to $(\chi, \theta_1, \theta_2, \ldots, \theta_{n-1})$; (1 – 1 transformation) by

$$\left.\begin{aligned} x_1 &= \chi\cos\theta_1\cos\theta_2 \ldots \cos\theta_{n-1} \\ x_2 &= \chi\cos\theta_1\cos\theta_2 \ldots \cos\theta_{n-2}\sin\theta_{n-1} \\ x_3 &= \chi\cos\theta_1\cos\theta_2 \ldots \cos\theta_{n-3}\sin\theta_{n-2} \\ &\vdots \\ &\vdots \\ x_j &= \chi\cos\theta_1\cos\theta_2 \ldots \cos\theta_{n-j}\sin\theta_{n-j+1} \\ &\vdots \\ &\vdots \\ x_n &= \chi\sin\theta_1 \end{aligned}\right\} \quad \ldots(13\cdot3$$

where $\chi > 0$, $-\pi < \theta_1 < \pi$ and $-\pi/2 < \theta_i < \pi/2$ for $i = 2, 3, \ldots (n-1)$.

Then $$x_1^2 + x_2^2 + \ldots + x_n^2 = \chi^2$$

and $$|J| = \chi^{n-1} \cos^{n-2}\theta_1 \cos^{n-3}\theta_2 \ldots \cos\theta_{n-2}$$

(*c.f.* Advanced Theory of Statistics Vol 1, by Kendall and Stuart.)

The joint distribution of $X_1, X_2, \ldots, X_n$ *viz.*,

$$dF(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp(-\Sigma x_i^2/2) \prod_{i=1}^{n} dx_i$$

transforms to

$$dG(\chi, \theta_1, \theta_2, \ldots, \theta_{n-1}) = \exp(-\chi^2/2)\, \chi^{n-1} \cos^{n-2}\theta_1 \cos^{n-3}\theta_2 \ldots \cos\theta_{n-2}\, d\chi d\theta_1 d\theta_2 \ldots d\theta_{n-1}$$

Integrating over $\theta_1, \theta_2, \ldots \theta_{n-1}$, we get the distribution of χ^2 as

$$dP(\chi^2) = k \exp(-\chi^2/2)\, (\chi^2)^{(n/2)-1}\, d\chi^2,\ 0 \le \chi^2 < \infty$$

The constant k is determined from the fact that the total probability is unity, *i.e.*,

$$\int_0^\infty dP(\chi^2) = 1 \Rightarrow k \int_0^\infty \exp(-\chi^2/2)\, (\chi^2)^{\frac{n}{2}-1}\, d\chi^2 = 1$$

$$\Rightarrow \qquad k = \frac{1}{2^{n/2}\,\Gamma(n/2)}$$

$$\therefore \qquad dP(\chi^2) = \frac{1}{2^{n/2}\,\Gamma(n/2)} \exp(-\chi^2/2)\, (\chi^2)^{\frac{n}{2}-1},\ 0 \le \chi^2 < \infty$$

Hence $\dfrac{\chi^2}{2} = \dfrac{1}{2}\sum_{i=1}^{n} X_i^2$ is a $\gamma(n/2)$ variate.

$\Rightarrow \qquad \chi^2 = \sum_{i=1}^{n} X_i^2$ is a chi-square variate with n degrees of freedom (d.f.) and (13·2) gives p.d.f. of chi-square distribution with n d.f.

Remarks 1. If X_i ; $i = 1, 2, \ldots, n$ are n independent normal variates with mean μ_i and S.D. σ_i, then $\sum_{i=1}^{n} \left(\dfrac{X_i - \mu_i}{\sigma_i}\right)^2$ is a χ^2-variate with n d.f.

2. In random sampling from a normal population with mean μ and S.D. σ, $\bar{x}$ is distributed normally about the mean μ with S.D. $\sigma/\sqrt{n}$.

$$\therefore \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$\Rightarrow \qquad \left[\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right]^2$ is a χ^2-variate with 1 d.f.

3. Normal distribution is a particular case of χ^2-distribution when $n = 1$, since for $n = 1$,

$$p(\chi^2) = \frac{1}{\sqrt{2}\,\Gamma(1/2)} \exp\,(-\chi^2/2)\,(\chi^2)^{\frac{1}{2}-1}\,d\chi^2,\ 0 \le \chi^2 < \infty$$
$$= \frac{1}{\sqrt{2\pi}} \exp\,(-\chi^2/2)\,d\chi,\ -\infty \le \chi < \infty$$

Thus χ is a standard normal variate.

13·3. M.G.F. of χ^2-distribution. Let $X \sim \chi^2_{(n)}$, then

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} f(x)dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty e^{tx}\,.\,e^{-x/2}\,x^{(n/2)-1}\,dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty \exp\left[-\left(\frac{1-2t}{2}\right)x\right].\,x^{(n/2)-1}\,dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)}\,\frac{\Gamma(n/2)}{[(1-2t)/2]^{n/2}} \qquad \text{[Using Gamma Integral]}$$

$$= (1-2t)^{-n/2},\ |2t| < 1 \qquad \ldots(13\cdot4)$$

which is the required m.g.f. of a χ^2-variate with n d.f.

Remarks 1. Using Binomial expansion for negative index, we get from (13·4) if $|t| < \frac{1}{2}$.

$$M(t) = 1 + \frac{n}{2}(2t) + \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)}{2\,!}(2t)^2 + \ldots$$
$$+ \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}+2\right)\ldots\left(\frac{n}{2}+r-1\right)}{r\,!}(2t)^r + \ldots$$

$$\mu_r' = \text{Coefficient of } \frac{t^r}{r\,!} \text{ in the expansion of } M(t)$$

$$= 2^r \frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}+2\right)\ldots\left(\frac{n}{2}+r-1\right)$$

$$= n(n+2)(n+4)\ldots(n+2r-2) \qquad \ldots(13\cdot4a)$$

Remark. If n is even so that $n/2$ is a positive integer, then

$$\mu_r' = 2^r\,\Gamma[(n/2)+r]/\Gamma(n/2) \qquad \ldots(13\cdot4b)$$

13·3·1. Cumulant Generating Function of χ^2-distribution. If $X \sim \chi^2_{(n)}$, then

$$K_{X^2}(t) = \log M_X(t) = -\frac{n}{2}\log\,(1-2t)$$

$$= \frac{n}{2}\left[2t + \frac{(2t)^2}{2} + \frac{(2t)^3}{3} + \frac{(2t)^4}{4} + \ldots\right]$$

$\therefore \quad \kappa_1 = \text{Coefficient of } t \text{ in } K(t) = n$

$\kappa_2 = \text{Coefficient of } \frac{t^2}{2\,!} \text{ in } K(t) = 2n$

$\kappa_3 = \text{Coefficient of } \frac{t^3}{3\,!} \text{ in } K(t) = 8n$

$\kappa_4 = \text{Coefficient of } \frac{t^4}{4\,!} \text{ in } K(t) = 48n$

In general,

$$\kappa_r = \text{Coefficient of } \frac{t^r}{r\,!} \text{ in } K(t) = n\, 2^{r-1}(r-1)\,! \qquad \ldots(13\cdot4c)$$

Hence

$$\left.\begin{array}{l} \text{Mean} = \kappa_1 = n, \ \text{Variance} = \mu_2 = \kappa_2 = 2n \\ \mu_3 = \kappa_3 = 8n, \ \mu_4 = \kappa_4 + 3\kappa_2^2 = 48n + 12n^2 \\ \beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{8}{n} \quad \text{and} \quad \beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{12}{n} + 3 \end{array}\right\} \qquad \ldots(13\cdot4d)$$

13·3·2. Limiting Form of χ^2 Distribution for Large Degrees of Freedom. If $X \sim \chi^2_{(n)}$, then $M_X(t) = (1-2t)^{-n/2}, \ |t| < \frac{1}{2}$

The m.g.f. of standard χ^2-variate Z is given by

$$M_{\frac{X-\mu}{\sigma}}(t) = e^{-\mu t/\sigma} M_X(t/\sigma) \qquad [\mu = n, \sigma^2 = 2n]$$

or

$$M_Z(t) = e^{-\mu t/\sigma}(1 - 2t/\sigma)^{-n/2}$$

$$= e^{-nt/\sqrt{2n}}\left(1 - \frac{2t}{\sqrt{2n}}\right)^{-n/2}$$

$$\therefore \quad K_Z(t) = \log M_Z(t) = -t\sqrt{\frac{n}{2}} - \frac{n}{2}\log\left(1 - t\sqrt{\frac{2}{n}}\right)$$

$$= -t\sqrt{\frac{n}{2}} + \frac{n}{2}\left[t\cdot\sqrt{\frac{2}{n}} + \frac{t^2}{2}\cdot\frac{2}{n} + \frac{t^3}{3}\left(\frac{2}{n}\right)^{3/2} + \ldots\right]$$

$$= -t\sqrt{\frac{n}{2}} + t\cdot\sqrt{\frac{n}{2}} + \frac{t^2}{2} + 0(n^{-1/2})$$

$$= \frac{t^2}{2} + 0(n^{-1/2}),$$

where $0(n^{-1/2})$ are terms containing $n^{1/2}$ and higher powers of n in the denominator.

$$\therefore \quad \lim_{n\to\infty} K_Z(t) = \frac{t^2}{2} \quad \Rightarrow \quad M_Z(t) = e^{t^2/2}, \text{ as } n \to \infty,$$

which is the m.g.f. of a standard normal variate. Hence by uniqueness theorem of m.g.f. Z is asymptotically normal. In other words, standard χ^2 variate tends to standard normal variate as $n \to \infty$. Thus, χ^2-distribution tends to normal distribution for large d.f.

In practice for $n \geq 30$, the χ^2-approximation to normal distribution is fairly good. So whenever $n \geq 30$, we use the normal probability tables for testing the significance of the value of χ^2. That is why in the tables given in the Appendix, the significant values of χ^2 have been tabulated till $n = 30$ only.

Remark. For the distribution of χ^2-variate for large values of n, see Example 13·7 and also Remark 2 to § 13·7·1.

13·3·3. Characteristic Function of χ^2-distribution.

If $X \sim \chi^2_{(n)}$, then

$$\phi_X(t) = E\{\exp(itX)\} = \int_0^\infty \exp(itx) f(x)\, dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty \exp\left\{-\left(\frac{1-2it}{2}\right)x\right\}(x)^{\frac{n}{2}-1}\, dx$$

$$= (1-2it)^{-n/2} \qquad \ldots(13·4e)$$

13·3·4. Mode and skewness of χ^2-distribution.

Let $X \sim \chi^2_{(n)}$, so that

$$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-x/2}\, x^{(n/2)-1},\ 0 \leq x < \infty \qquad \ldots(*)$$

Mode of the distribution is the solution of

$$f'(x) = 0 \quad \text{and} \quad f''(x) < 0$$

Logarithmic differentiation w.r.t. x in (*) gives :

$$\frac{f'(x)}{f(x)} = 0 - \frac{1}{2} + \left(\frac{n}{2} - 1\right) \cdot \frac{1}{x} = \frac{n-2-x}{2x} \qquad \ldots(13·5)$$

Since $f(x) \neq 0$, $f'(x) = 0 \Rightarrow x = n-2$

It can be easily seen that at the point, $x = (n-2), f''(x) < 0$.

Hence mode of the chi-square distribution with n d.f. is $(n-2)$.

Also Karl Pearson's coefficient of skewness is given by

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{n-(n-2)}{\sqrt{2n}} = \sqrt{\frac{2}{n}} \qquad \ldots(13·6)$$

Since Pearson's coefficient of skewness is greater than zero for $n \geq 1$, the χ^2-distribution is positively skewed. Further since skewness is inversely proportional to the square root of d.f., it rapidly tends to symmetry as the d.f. increases and consequently as $n \to \infty$, the chi-square distribution tends to normal distribution.

13·3·5. Additive Property of χ^2-variates. *The sum of independent chi-square variates is also a χ^2-variate. More precisely, if X_i, $(i = 1, 2, \ldots, k)$ are* *independent χ^2-variates with n_i d.f. respectively, then the sum $\sum_{i=1}^{k} X_i$ is also a chi-square variate with $\sum_{i=1}^{k} n_i$ d.f.*

Proof. We have

$$M_{X_i}(t) = (1-2t)^{-n_i/2}\ ;\ i = 1, 2, \ldots, k.$$

The *m.g.f.* of the sum $\sum_{i=1}^{k} X_i$ is given by

$$M_{\Sigma X_i}(t) = M_{X_1}(t)\, M_{X_2}(t) \ldots M_{X_k}(t) \qquad [\because X_i\text{'s are independent}]$$
$$= (1-2t)^{-n_1/2}(1-2t)^{-n_2/2} \ldots (1-2t)^{-n_k/2}$$
$$= (1-2t)^{-(n_1+n_2+\ldots+n_k)/2}$$

which is the *m.g.f.* of a χ^2-variate with $(n_1+n_2+\ldots+n_k)$ *d.f.* Hence by uniqueness theorem of *m.g.f.*'s, $\sum_{i=1}^{k} X_i$ is a χ^2-variate with $\sum_{i=1}^{k} n_i$ *d.f.*

Remarks 1. Converse is also true, *i.e.*, if X_i ; $i = 1, 2, \ldots, k$ are χ^2-variates with n_i ; $i = 1, 2, \ldots, k$ *d.f.* respectively and if $\sum_{i=1}^{k} X_i$ is a χ^2-variate with $\sum_{i=1}^{k} n_i$ *d.f.*, then X_i's are independent.

2. Another useful version of the converse is as follows :

If X and Y are independent non-negative variates such that $X + Y$ follows chi-square distribution with $n_1 + n_2$ *d.f.* and if one of them, say, X is a χ^2-variate with n_1 *d.f.* then the other, *viz.*, Y, is a χ^2-variate with n_2 *d.f.*

Proof. Since X and Y are independent variates, we have

$$M_{X+Y}(t) = M_X(t)\, M_Y(t)$$
$$\Rightarrow \quad (1-2t)^{-(n_1+n_2)/2} = (1-2t)^{-n_1/2} \cdot M_Y(t)$$
$$\left[\because X+Y \sim \chi^2_{(n_1+n_2)} \text{ and } X \sim \chi^2_{(n_1)}\right]$$
$$\Rightarrow \quad M_Y(t) = (1-2t)^{-n_2/2}$$

which is the m.g.f. of χ^2-variate with n_2 *d.f.* Hence by uniqueness theorem of m.g.f.'s, $Y \sim \chi^2_{(n_2)}$

3. Still another form of the above theorem is "*Cochran theorem*" which is as follows :

Let $X_1, X_2, \ldots, X_n$ be independently distributed as standard normal variates, *i.e.*, $N(0, 1)$. Let

$$\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \ldots + Q_k,$$

where each Q_i is a sum of squares of linear combinations of $X_1, X_2, \ldots, X_n$ with n_i degrees of freedom. Then if $n_1 + n_2 + \ldots + n_k = n$, the quantities $Q_1, Q_2 \ldots, Q_k$ are independent χ^2-variates with $n_1, n_2, \ldots, n_k$ d.f. respectively.

13·4. Chi-square Probability Curve. We get from (13·5)

$$f'(x) = \left[\frac{n-2-x}{2x}\right] f(x). \qquad \ldots(13\cdot7)$$

Since $x > 0$ and $f(x)$ being p.d.f. is always non-negative, we get from (13·7):

$$f'(x) < 0 \quad \text{if} \quad (n-2) \le 0,$$

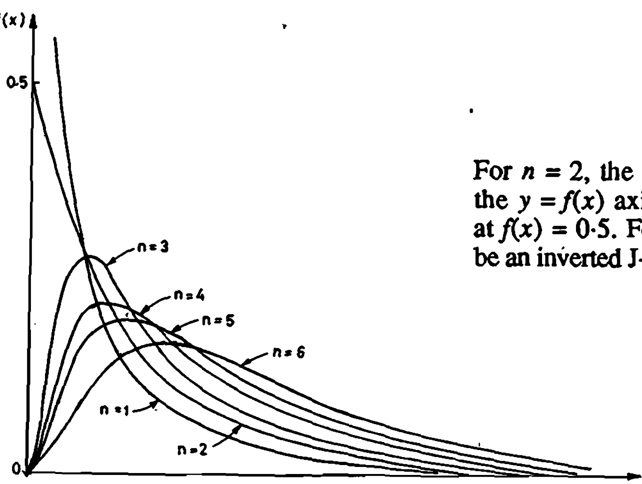
for all values of x. Thus the χ^2-probability curve for 1 and 2 degrees of freedom is monotonically decreasing.

When $n > 2$,

$$f'(x) = \begin{cases} > 0, \text{ if } x < (n-2) \\ = 0, \text{ if } x = n-2 \\ < 0, \text{ if } x > (n-2) \end{cases}$$

This implies that for $n > 2$, $f(x)$ is monotonically increasing for $0 < x < (n-2)$ and monotonically decreasing for $(n-2) < x < \infty$, while at $x = n-2$, it attains the maximum value.

For $n \geq 1$, as x increases, $f(x)$ decreases rapidly and finally tends to zero as $x \to \infty$. Thus for $n > 1$, the χ^2-probability curve is positively skewed [*c.f.* (13·6)] towards higher values of x. Moreover, x-axis is an asymptote to the curve. The shape of the curve for $n = 1, 2, 3, \ldots, 6$ is given below.



For $n = 2$, the curve will meet the $y = f(x)$ axis at $x = 0$, *i.e.*, at $f(x) = 0\cdot5$. For $n = 1$, it will be an inverted J-shaped curve.

PROBABILITY CURVE OF CHI-SQUARE DISTRIBUTION

Theorem 13·1. *If χ_1^2 and χ_2^2 are two independent χ^2-variates with n_1 and n_2 d.f. respectively, then*

$$\frac{\chi_1^2}{\chi_2^2} \text{ is a } \beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right) \text{variate.}$$

(Gauhati Univ. M.Sc., 1992)

Proof. Since χ_1^2 and χ_2^2 are independent χ^2-variates with n_1 and n_2 d.f. respectively, their joint probability differential is given by the compound probability theorem as

$$dP(\chi_1^2, \chi_2^2) = dP_1(\chi_1^2)\, dP_2(\chi_2^2)$$

$$= \left[\frac{1}{2^{n_1/2}\Gamma(n_1/2)} \exp(-\chi_1^2/2)(\chi_1^2)^{(n_1/2)-1} d\chi_1^2\right]$$

$$\times \left[\frac{1}{2^{n_2/2}\Gamma(n_2/2)} \exp(-\chi_2^2/2)(\chi_2^2)^{(n_2/2)-1} d\chi_2^2\right]$$

$$= \frac{1}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} \exp\{-(\chi_1^2+\chi_2^2)/2\}$$

$$\times (\chi_1^2)^{\frac{n_1}{2}-1} (\chi_2^2)^{\frac{n_2}{2}-1} d\chi_1^2 d\chi_2^2,\ 0 \le (\chi_1^2, \chi_2^2) < \infty$$

Let us make the transformation :

$$u = \chi_1^2/\chi_2^2 \quad \text{and} \quad v = \chi_2^2$$

so that $\chi_1^2 = uv$ and $\chi_2^2 = v$

Jacobian of transformation J is given by

$$J = \frac{\partial(\chi_1^2, \chi_2^2)}{\partial(u, v)} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$$

Thus the joint distribution of random variables U and V becomes

$$dG(u, v) = \frac{1}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} \exp\{-(1+u)v/2\}$$

$$\times (uv)^{\frac{n_1}{2}-1} v^{\frac{n_2}{2}-1} |J|\, du\, dv,$$

$$= \frac{1}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} \exp\{-(1+u)v/2\}$$

$$\times u^{\frac{n_1}{2}-1} v^{\frac{n_1+n_2}{2}-1} du\, dv,\ 0 \le (u, v) < \infty$$

Integrating w.r.t. v over the range 0 to ∞, we get the marginal distribution of U as : $dG_1(u) = \int_0^\infty dG(u, v)$

$$= \frac{1}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} u^{(n_1/2)-1} du$$

$$\times \int_0^\infty \exp\left\{-\left(\frac{1+u}{2}\right)v\right\} v^{[(n_1+n_2)/2]-1} dv$$

$$= \frac{u^{(n_1/2)-1}}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot \frac{\Gamma\{(n_1+n_2)/2\}}{[(1+u)/2]^{(n_1+n_2)/2}}\,du$$

$$= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{u^{(n_1/2)-1}}{[+u]^{(n_1+n_2)/2}}\,du,\ 0 \le u < \infty$$

Hence $U = \dfrac{\chi_1^2}{\chi_2^2}$ is a $\beta_2\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)$ variate.

Theorem 13·2. *If χ_1^2 and χ_2^2 are independent χ^2-variates with n_1 and n_2 d.f. respectively, then*

$$U = \frac{\chi_1^2}{\chi_1^2 + \chi_2^2} \quad \textit{and} \quad V = \chi_1^2 + \chi_2^2$$

are independently distributed, U as a $\beta_1\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)$ variate and V as a χ^2 variate with (n_1+n_2) d.f.

Proof. As the Theorem 13·1, we have

$$dP(\chi_1^2, \chi_2^2) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\ \{-(\chi_1^2+\chi_2^2)/2\}$$
$$\times (\chi_1^2)^{(n_1/2)-1} (\chi_2^2)^{(n_2/2)-1}\,d\chi_1^2\,d\chi_2^2,\ 0 \le (\chi_1^2, \chi_2^2) < \infty$$

Let us transform to u and v defined as follows:

$$u = \frac{\chi_1^2}{\chi_1^2+\chi_2^2} \quad \text{and} \quad v = \chi_1^2 + \chi_2^2$$

so that $\chi_1^2 = uv$ and $\chi_2^2 = v - \chi_1^2 = (1-u)v$

As χ_1^2 and χ_2^2 both range from 0 to ∞, u ranges from 0 to 1 and v from 0 to ∞.

Jacobian of transformation J is

$$J = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

$$\therefore\ dG(u, v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\ (-v/2)\,(uv)^{(n_1/2)-1}$$
$$\times \left[(1-u)\,v\right]^{(n_2/2)-1}\,|J|\,du\,dv$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)}\,u^{(n_1/2)-1}\,(1-u)^{(n_2/2)-1}$$
$$\times \exp\ (-v/2).\ v^{\{(n_1+n_2)/2\}-1}\,du\,dv$$

$$= \left[\frac{\Gamma\{(n_1+n_2)/2\}}{\Gamma(n_1/2)\Gamma(n_2/2)}\,u^{(n_1/2)-1}\,(1-u)^{(n_2/2)-1}\,du\right]$$
$$\times \left[\frac{1}{2^{(n_1+n_2)/2}\,\Gamma\{(n_1+n_2)/2\}} \exp\ (-v/2)\,v^{\{(n_1+n_2)/2\}-1}\,dv\right]$$

Since the joint probability differential of U and V is the product of their respective probability differentials, U and V are independently distributed, with

$$dG_1(u) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} u^{(n_1/2)-1} (1-u)^{(n_2/2)-1}\, du, \quad 0 \le u \le 1$$

and

$$dG_2(v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma\{(n_1+n_2)/2\}} \exp(-v/2)\, v^{\{(n_1+n_2)/2\}-1}\, dv,$$
$$0 \le v < \infty$$

i.e., U as a $\beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate and V as a χ^2-variate with $(n_1 + n_2)$ d.f

Remark. The results in Theorems 13·1 and 13·2 can be summarised as follows :

If $X \sim \chi^2_{(n_1)}$ and $Y \sim \chi^2_{(n_2)}$ are independent chi-square variates then :

(*i*) $X + Y \sim \chi^2_{(n_1+n_2)}$ *i.e.*, the sum of two independent chi-square variates is also a chi-square variate.

(*ii*) $\frac{X}{Y} \sim \beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ *i.e.*, the ratio of two independent chi-square variates is a β_2-variate.

(*iii*) $\frac{X}{X+Y} \sim \beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$

Theorem 13·3. *In a random and large sample,*

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(n_i - np_i)^2}{np_i}\right], \qquad \ldots(13\cdot8)$$

follows chi-square distribution approximately with (k − 1) degrees of freedom, where n_i is the observed frequency and np_i is the corresponding expected frequency of the ith class, (i = 1, 2, ..., k), $\sum_{i=1}^{k} n_i = n$.

Proof. Let us consider a random sample of size n, whose members are distributed at random in k classes or cells. Let p_i be the probability that sample observation will fall in the i th cell, (i = 1, 2, ..., k). Then the probability P of there being n_i members in the ith cell, (i = 1, 2, ..., k) respectively is given by the multinomial probability law, by the expression

$$P = \frac{n!}{n_1!\; n_2! \;\ldots\; n_k!}\, p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$

where $\sum_{i=1}^{k} n_i = n$ and $\sum_{i=1}^{k} p_i = 1$.

If n is sufficiently large so that n_i, $(i = 1, 2, \ldots, k)$ are not small then using Stirling's approximation to factorials for large n, *viz.*,

$$\lim_{n \to \infty} (n\,!) \approx \sqrt{2\pi}\, e^{-n}\, n^{\,n + \frac{1}{2}}, \text{ we get}$$

$$P \approx \frac{\sqrt{2\pi}\, e^{-n}\, n^{n+\frac{1}{2}}}{(\sqrt{2\pi})^k e^{-(n_1 + n_2 + \ldots + n_k)}} \times \frac{p_1^{\,n_1}\, p_2^{\,n_2} \ldots p_k^{\,n_k}}{n_1^{\,n_1 + \frac{1}{2}}\, n_2^{\,n_2 + \frac{1}{2}} \ldots n_k^{\,n_k + \frac{1}{2}}}$$

$$\approx \frac{e^{-n}\, n^{n+\frac{1}{2}} \left(\dfrac{np_1}{n_1}\right)^{n_1 + \frac{1}{2}} \left(\dfrac{np_2}{n_2}\right)^{n_2 + \frac{1}{2}} \cdots \left(\dfrac{np_k}{n_k}\right)^{n_k + \frac{1}{2}}}{(\sqrt{2\pi})^{k-1}\, e^{-n}\, n^{\,n_1 + n_2 + \ldots + n_k + (k/2)}\, (p_1 p_2 \ldots p_k)^{1/2}}$$

$$\approx C \prod_{i=1}^{k} \left(\frac{np_i}{n_i}\right)^{n_i + \frac{1}{2}}$$

where $C = \dfrac{1}{(2\pi)^{(k-1)/2}\, n^{(k-1)/2}\, (p_1 p_2 \ldots p_k)^{1/2}}$,

is a constant independent of n_i's.

$$\therefore \qquad \log P \approx \log C + \sum_{i=1}^{k} (n_i + \tfrac{1}{2}) \log \left(\frac{np_i}{n_i}\right)$$

$$\Rightarrow \qquad \log (P/C) \approx \sum_{i=1}^{k} (n_i + \tfrac{1}{2}) \log \left(\frac{\lambda_i}{n_i}\right), \qquad \ldots(*)$$

where $\lambda_i = np_i$ is the expected frequency for the ith cell, *i.e.*,

$$E(n_i) = np_i = \lambda_i,\ (i = 1, 2, \ldots, k).$$

Let us define

$$\xi_i = \frac{n_i - \lambda_i}{\sqrt{\lambda_i}},$$

so that $n_i - \lambda_i = \xi_i \sqrt{\lambda_i} \quad \Rightarrow \quad n_i = \lambda_i + \xi_i \sqrt{\lambda_i}$ $\qquad \ldots(**)$

Substituting in (*), we get

$$\log (P/C) \approx \sum_{i=1}^{k} (\lambda_i + \xi_i \sqrt{\lambda_i} + \tfrac{1}{2}) \log \left[\frac{\lambda_i}{\lambda_i + \xi_i \sqrt{\lambda_i}}\right]$$

$$= \sum_{i=1}^{k} (\lambda_i + \xi_i \sqrt{\lambda_i} + \tfrac{1}{2}) \log [1/\{1 + \xi_i/\sqrt{\lambda_i}\}]$$

$$= -\sum_{i=1}^{k} (\lambda_i + \xi_i \sqrt{\lambda_i} + \tfrac{1}{2}) \log \{1 + (\xi_i/\sqrt{\lambda_i})\}$$

If we assume that ξ_i is small compared with λ_i, the expansion of $\log 1 + (\xi_i/\sqrt{\lambda_i}\}$ in ascending powers of $\xi_i/\sqrt{\lambda_i}$ is valid.

$$\therefore \quad \log P/C \approx -\sum_{i=1}^{k} (\lambda_i + \xi_i\sqrt{\lambda_i} + \tfrac{1}{2}\lambda_i)\left[\frac{\xi_i}{\sqrt{\lambda_i}} - \frac{1}{2}\frac{\xi_i^2}{\lambda_i} + O\,(1/\lambda_i^{3/2})\right]$$

$$\approx -\sum_{i=1}^{k} [\xi_i\sqrt{\lambda_i} - \tfrac{1}{2}\xi_i^2 + \xi_i^2 + O(\lambda_i^{-1/2})],$$

neglecting higher powers of $\xi_i/\sqrt{\lambda_i}$ if ξ_i is small compred with λ_i.
Since n is large, so is $\lambda_i = np_i$. Hence $O(\lambda_i^{-1/2}) \to 0$ for large n.

$$\text{Also} \quad \sum_{i=1}^{k} \xi_i\sqrt{\lambda_i} = \sum_{i=1}^{k} (n_i - \lambda_i) = \sum_{i=1}^{k} n_i - \sum_{i=1}^{k} \lambda_i$$

$$= \sum_{i=1}^{k} n_i - n\sum_{i=1}^{k} p_i = n - n = 0 \qquad (\because \Sigma n_i = n,\ \Sigma p_i = 1)$$

$$\therefore \quad \log (P/C) \approx -\left[\sum_{i=1}^{k} \xi_i\sqrt{\lambda_i} + \tfrac{1}{2}\sum_{i=1}^{k} \xi_i^2 + O(\lambda_i^{-1/2})\right] \approx -\tfrac{1}{2}\sum_{i=1}^{k} \xi_i^2$$

$$\Rightarrow \qquad P \approx C \exp\left(-\tfrac{1}{2}\sum_{i=1}^{k} \xi_i^2\right),$$

which shows that ξ_i, $(i = 1, 2, \ldots, k)$ are distributed as independent standard normal variates.

$$\text{Hence} \qquad \sum_{i=1}^{k} \xi_i^2 = \sum_{i=1}^{k} \left[\frac{(n_i - \lambda_i)^2}{\lambda_i}\right],$$

being the sum of the squares of k independent standard normal variates is a χ^2-variate with $(k-1)$ d.f., one d.f. being lost because of the linear constraint

$$\sum_{i=1}^{k} \xi_i\sqrt{\lambda_i} = \Sigma(n_i - \lambda_i) = 0 \Rightarrow \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \lambda_i \qquad \ldots(***)$$

Remarks 1. If O_i and E_i $(i = 1, 2, \ldots, k)$, be a set of observed and expected frequencies, then

$$\chi^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right], \left(\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i\right) \qquad \ldots(13\cdot8a)$$

follows chi-square distribution with $(k-1)$ d.f

Another convenient form of this formula is as follows :

$$\chi^2 = \sum_{i=1}^{k}\left(\frac{O_i^2 + E_i^2 - 2O_iE_i}{E_i}\right) = \sum_{i=1}^{k}\left(\frac{O_i^2}{E_i} + E_i - 2O_i\right)$$

$$= \sum_{i=1}^{k}\left(\frac{O_i^2}{E_i}\right) + \sum_{i=1}^{k} E_i - 2\sum_{i=1}^{k} O_i$$

$$= \sum_{i=1}^{k} \left(\frac{O_i^2}{E_i} \right) - N, \qquad \ldots(13{\cdot}8b)$$

where $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = N$ (say), is the total frequency.

2. Conditions for the Validity of χ^2-test. χ^2-test is an approximate test for large values of n. For the validity of chi-square test of '*goodness of fit*' between theory and experiment, the following conditions must be satisfied :

(*i*) The sample observations should be independent.

(*ii*) Constraints on the cell frequencies, if any, should be linear, *e.g.*, $\sum n_i = \sum \lambda_i$ or $\sum O_i = \sum E_i$.

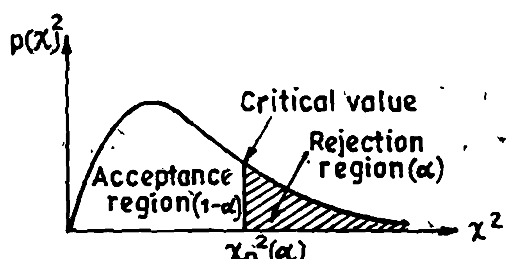
(*iii*) N, the total frequency should be reasonably large, say, greater than 50.

(*iv*) No theoretical cell frequency should be less than 5. (The chi square distribution is essentially a continuous distribution but it cannot maintain its character of continuity if cell frequency is less than 5). If any theoretical cell frequency is less than 5, then for the application of χ^2-test, it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.

3. It may be noted that the χ^2-test depends only on the set of observed and expected frequencies and on degrees of freedom (*d.f.*). It does not make any assumptions regarding the parent population from which the observations are taken. Since χ^2 defined in (13·8) does not involve any population parameters, it is termed as a statistic and the test is known as *Non-Parametric Test* or *Distribution-Free Test.*

4. Critical Values. Let $\chi_n^2(\alpha)$ denote the value of chi-square for n *d.f.* such that the area to the right of this point is α, *i.e.*,

$$P[\chi^2 > \chi_n^2(\alpha)] = \alpha \qquad \ldots(13{\cdot}8c)$$



The value $\chi_n^2(\alpha)$ defined in (13·8*c*) is known as the *upper* (*right-tailed*) *α-point or Critical Value or Significant Value of chi-square for n d.f.* and has been tabulated for different values of n and α in Table VI in the Appendix at the end of the book. From these tables we observe that the critical values of χ^2 increase as n (d.f.) increases and level of significance (α) decreases.

The critical values for left-tailed test or two tailed tests can be obtained from the above table, as discussed in Remark 1 to § 16·7·4.

13·6. Linear Transformation. Let us suppose that the given set of variables $X' = (x_1, x_2, \ldots x_n)$ is transformed to a new set of variables $Y' = (y_1, y_2, \ldots, y_n)$ by means of the linear transformation :

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ &\vdots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{aligned}\right] \quad \ldots(13\cdot9)$$

i.e., $\quad y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \; ; i = 1, 2, \ldots, n$

In matrix notation, this system of linear equations can be expressed symbolically as

$$Y = AX \qquad \ldots(13\cdot10)$$

where $\quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \; A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$

From matrix theory, we know that the system (13·10) has a unique solution iff $|A| \neq 0$. In other words, we can express X uniquely in terms Y if A is non-singular and the solution is given by

$$X = A^{-1}Y \qquad \ldots(13\cdot10a)$$

where A^{-1} is the inverse of the square matrix A.

The linear transformation defined in (13·9) or (13·10) is said to be *orthogonal* if

$$X'X = Y'Y \qquad \ldots(13\cdot11)$$

$$\Rightarrow \quad X'X = (AX)'AX = X'(A'A)X$$

$$\Rightarrow \quad A'A = I_n \qquad \ldots(13\cdot11a)$$

$\Rightarrow$ A is an orthogonal matrix.

More elaborately

$$X'X = Y'Y$$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n)^2, \qquad \ldots(*)$$

for every set of variables, $(x_1, x_2, \ldots, x_n)$.

If we write $\delta_{ij} = \sum_{k=1}^{n} a_{ik}\, a_{kj}, \; (i, j = 1, 2, \ldots, n)$,

then (*) implies that δ_{ij} is a Kronecker delta so that

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \qquad \ldots(13\cdot11b)$$

whence it follows that A is an orthogonal matrix.

Linear Orthogonal Transformation. Def. : *A linear transformation* $Y = AX$, *is said to be otthogonal if* A *is an orthogonal matrix.*

Remarks 1. It is very easy to verify the equivalence of the following two definitions of an orthogonal matrix.

Def. 1. A square matrix A $(n \times n)$ is said to be orthogonal if $A'A = AA' = I_n$.

Def. 2. A square matrix A is said to be orthogonal if the transformation $Y = AX$ transforms $X'X$ to $Y'Y$.

2. If $Y = AX$ is an orthogonal transformation, then $Y'Y = X'X$ and $A'A = AA' = I_n$.

Theorem 13·4. (*Fisher's Lemma*). *If* X_i, $(i = 1, 2, \ldots, n)$ *are independent* $N(0, \sigma^2)$ *and they are transformed to a new set of variables* Y_i, $(i = 1, 2, \ldots, n)$, *by means of a linear orthogonal transformation, then* Y_i, $(i = 1, 2, \ldots, n)$ *are also independent* $N(0, \sigma^2)$.

Proof. Let the linear orthogonal transformation be

$$Y = AX \quad \text{so that} \quad Y'Y = X'X \text{ and } A'A = I_n$$

Since X_i, $(i = 1, 2, \ldots, n)$ are independent $N(0, \sigma^2)$, their joint density function is given by

$$f(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left(-\sum_{i=1}^{n} x_i^2/2\sigma^2\right), -\infty < (x_1, x_2, \ldots, x_n) < \infty$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-X'X/2\sigma^2\right)$$

The joint density of $(Y_1, Y_2, \ldots, Y_n)$ becomes

$$g(y_1, y_2, \ldots, y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-Y'Y/\sigma^2\right) |J|$$

where

$$\frac{1}{J} = \frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(x_1, x_2, \ldots, x_n)} = |A|$$

$$\begin{aligned}
\text{Now} \quad & A'A = I_n \\
\Rightarrow \quad & |A'A| = |I_n| = 1 \\
\Rightarrow \quad & |A'||A| = 1 \\
\Rightarrow \quad & |A|^2 = 1 \qquad (\because |A'| = |A|) \\
\Rightarrow \quad & |A| = \pm 1 \\
\therefore \quad & |J| = |\pm 1| = 1
\end{aligned}$$

$$\therefore \quad g(y_1, y_2, \ldots, y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-Y'Y/2\sigma^2\right)$$

$$= \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-y_i^2/2\sigma^2\right)\right]$$

Hence Y_i, $(i = 1, 2, \ldots, n)$ are independent $N(0, \sigma^2)$.

Theorem 13·5. *Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. Then*

(*i*) $\overline{X} \sim N(\mu, \sigma^2/n)$,

(*ii*) $\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma}\right)^2$ *is a χ^2-variate with $(n-1)$ d.f., and*

(*iii*) $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $\frac{ns^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma}\right)^2$ *are independently distributed.*

[**Delhi Univ. B.Sc. (Maths Hons.) 1987; Sardar Patel Univ. B.Sc. 1992**]

Proof. The joint probability differential of $X_1, X_2, \ldots, X_n$ is given by

$$dP(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right] dx_1\, dx_2 \ldots dx_n\ ;$$

$$-\infty < (x_1, x_2, \ldots, x_n) < \infty$$

Let us transform to the variables Y_i, $(i = 1, 2, \ldots, n)$ by means of a linear orthogonal transformation $(\mathbf{Y} = \mathbf{AX})$ (*c.f.* § 13·6, page 13·16).

Let us choose in particular

$$a_{11} = a_{12} = \ldots = a_{1n} = 1/\sqrt{n}$$

$$\Rightarrow \qquad y_1 = \frac{1}{\sqrt{n}}(x_1 + x_2 + \ldots + x_n) = \sqrt{n}\ \bar{x} \qquad \ldots(*)$$

(It can be easily seen that the above choice of $a_{11}, a_{12}, \ldots, a_{1n}$ satisfies the condition of orthogonality; *viz.*, $\sum_{i=1}^{n} a_{ij}^2 = 1$).

Since the transformation is orthogonal, we have

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + y_1^2 \qquad \text{[From (*)]}$$

$$\Rightarrow \qquad \sum_{i=2}^{n} y_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \ldots(**)$$

Also $$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

$$= \sum_{i=2}^{n} y_i^2 + n(\bar{x} - \mu)^2 \qquad \text{[From (**)]}$$

As in Theorem 13·4, the Jacobian of transformation $J = \pm 1$.

Thus the joint density function of $X_1, X_2, \ldots, X_n$ transforms to

$$dG(y_1, y_2, \ldots, y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=2}^{n} y_i^2 + n(\bar{x}-\mu)^2\right\}\right]$$

$$\times |J|\, dy_1\, dy_2 \ldots dy_n$$

$$= \left[\frac{1}{\sqrt{2\pi}(\sigma/\sqrt{n})} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right\} d\bar{x}\right]$$

$$\times \left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n-1} \exp\left\{-\sum_{i=2}^{n} \frac{y_i^2}{2\sigma^2}\right\} dy_2\, dy_3 \ldots dy_n\right]$$

$$(\because dy_1 = \sqrt{n}\, d\bar{x})$$

Thus $\bar{X}$ and $\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 = ns^2$, (where s^2 is the sample variance), are independently distributed, which establishes part (*iii*) of the Theorem. Moreover $\bar{X} \sim N(\mu, \sigma^2/n)$ and Y_i, $(i = 1, 2, 3, \ldots, n)$ are independent $N(0, \sigma^2)$. Hence

$$\sum_{i=2}^{n} \frac{Y_i^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2,$$

being the sum of squares of $(n-1)$ independent standard normal variates is distributed as χ^2-variate with $(n-1)$ d.f.

Aliter. The alternative proof of the above Theorem is based on the use of *m.g.f.*'s and is given below.

We shall first prove that :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad X_i - \bar{X},\ i = 1, 2, \ldots n \qquad \ldots(i)$$

are independently distributed.

The joint m.g.f. of $\bar{X}$ and $(X_i - \bar{X})$ is given by :

$$M(t_1, t_2) = E\left[e^{t_1\bar{X} + t_2(X_i - \bar{X})}\right] = E\left[e^{(t_1 - t_2)\bar{X} + t_2 X_i}\right]$$

$$= E\left[\exp\left\{\frac{t_1 - t_2}{n} \cdot \sum_{i=1}^{n} X_i + t_2 X_i\right\}\right]$$

$$= E\left[\exp\left\{\left(\frac{t_1 - t_2}{n} + t_2\right)X_i + \frac{t_1 - t_2}{n} \sum_{\substack{j=1 \\ (j \neq i)}}^{n} X_j\right\}\right]$$

$$= E\left[\exp\left\{\left(\frac{t_1 - t_2}{n} + t_2\right)X_i\right\}\right] . E\left[\exp\left(\frac{t_1 - t_2}{n} \sum_{\substack{j=1 \\ (j \neq i)}}^{n} X_j\right)\right] \qquad ...(ii)$$

($\because X_1, X_2, \ldots, X_n$ are independent)

Now $U = \sum_{\substack{j=1 \\ (j \neq i)}}^{n} X_j$, being the sum of $(n - 1)$ i.i.d. $N(\mu, \sigma^2)$ variates is a $N\{(n-1)\mu, (n-1)\sigma^2\}$ variate.

$$\therefore \quad M_U(t) = \exp\left[t . (n-1)\mu + t^2 . (n-1)\sigma^2/2\right] \qquad ...(iii)$$

Hence

$$E\left[\exp\left\{\frac{t_1 - t_2}{n} \sum_{\substack{j=1 \\ (j \neq i)}}^{n} X_j\right\}\right] = E\left[\exp\left(\frac{t_1 - t_2}{n} . U\right)\right]$$

$$= M_U[(t_1 - t_2)/n]$$

$$= \exp\left[\left(\frac{t_1 - t_2}{n}\right)(n-1)\mu + \left(\frac{t_1 - t_2}{n}\right)^2 (n-1)\frac{\sigma^2}{2}\right]$$

[On using (iii)] ...(iv)

$$\text{and } E\left[\exp\left\{\left(\frac{t_1 - t_2}{n} + t_2\right)X_i\right\}\right] = M_{X_i}\left(\frac{t_1 - t_2}{n} + t_2\right)$$

$$= \exp\left[\left(\frac{t_1 - t_2}{n} + t_2\right)\mu + \left(\frac{t_1 - t_2}{n} + t_2\right)^2 \frac{\sigma^2}{2}\right]$$

[$\because X_i \sim N(\mu, \sigma^2)$] ...(v)

Substituting from (iv) and (v) in (ii), we get

$$M(t_1, t_2) = \exp\left[\left\{\left(\frac{t_1 - t_2}{n}\right)(n-1) + \left(\frac{t_1 - t_2}{n} + t_2\right)\right\}\mu\right]$$

$$\times \exp\left[\left\{\left(\frac{t_1 - t_2}{n}\right)^2 . (n-1) + \left(\frac{t_1 - t_2}{n} + t_2\right)^2\right\}\frac{\sigma^2}{2}\right]$$

$$= \exp\left[t_1 \mu + \frac{1}{2} t_1^2 \left(\frac{\sigma^2}{n}\right)\right] \times \exp\left[\frac{1}{2} t_2^2 \left(\frac{n-1}{n}\right)\sigma^2\right]$$

(On simplification)

$$= M(t_1) . M(t_2)$$

$\Rightarrow$ (a) $\overline{X}$ and $X_i - \overline{X}$; $i = 1, 2, \ldots, n$ are independently distributed ... (vi)

and (b) $\overline{X} \sim N(\mu, \sigma^2/n)$ ...(vii)

and $X_i - \overline{X} \sim N\left(0, \frac{n-1}{n}\sigma^2\right)$ ...(viii)

Since $\overline{X}$ and $X_i - \overline{X}$; $i = 1, 2, \ldots, n$ are independently distributed,

$$\overline{X} \quad \text{and} \quad s^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2, \qquad \ldots(viii\ a)$$

are independently distributed.

To derive the distribution of s^2, we note that :

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \overline{X} + \overline{X} - \mu)^2$$

$$= \sum_{i=1}^{n} (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2,$$

the product term vanishes since $\sum_{i=1}^{n} (X_i - \overline{X}) = 0$.

$$\therefore \quad \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} + \left[\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right]^2 \qquad \ldots(ix)$$

$$\Rightarrow \quad V = W + Z, \qquad \ldots(ix\ a)$$

where $V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$, being the sum of squares of n independent standard normal variates is a $\chi^2_{(n)}$ variate. Hence

$$M_V(t) = (1 - 2t)^{-n/2} \; ; \; |t| < 1/2, \qquad \ldots(x)$$

Also $\overline{X} \sim N(\mu, \sigma^2/n) \Rightarrow \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

$$\therefore \quad Z = \left[\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right]^2 \sim \chi^2_{(1)}$$

$$\Rightarrow \quad M_Z(t) = (1 - 2t)^{-1/2} \qquad \ldots(xi)$$

Further, since $\overline{X}$ and s^2 are independent, [see *viii (a)*], W and Z are independently distributed.

$$\therefore \quad M_V(t) = M_{W+Z}(t) = M_W(t) \cdot M_Z(t)$$

($\because$ W and Z are independent)

$$\Rightarrow \quad (1 - 2t)^{-n/2} = M_W(t) \cdot (1 - 2t)^{-1/2} \qquad \text{[From } (x) \text{ and } (xi)]$$

$$\Rightarrow \quad M_W(t) = (1 - 2t)^{-(n-1)/2}, \quad |t| < 1/2$$

which is the m.g.f. of χ^2-variate with $(n - 1)$ d.f. Hence by uniqueness theorem of m.g.f.

$$W = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n-1)} \qquad \ldots(xii)$$

Remarks 1. *p.d.f. of the sample variance* $s^2 = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

Since $$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 = ns^2/\sigma^2$$

is a χ^2-variate with $(n-1)$ d.f., we have

$$dP(ns^2/\sigma^2) = \frac{1}{2^{(n-1)/2}\,\Gamma[(n-1)/2]} \cdot e^{-ns^2/2\sigma^2} \left(\frac{ns^2}{\sigma^2}\right)^{\frac{n-1}{2}-1} d(ns^2/\sigma^2)$$

$$\Rightarrow \quad dP(s^2) = \frac{(n/2\sigma^2)^{(n-1)/2}}{\Gamma[(n-1)/2]} \cdot e^{-ns^2/2\sigma^2} (s^2)^{(n-3)/2}\, ds^2,\ 0 < s^2 < \infty.$$

2. We have $$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\therefore \quad E\left(\frac{ns^2}{\sigma^2}\right) = n-1$$

$$\Rightarrow \quad \frac{n}{\sigma^2} E(s^2) = (n-1)$$

$$\Rightarrow \quad E(s^2) = \left(\frac{n-1}{n}\right)\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2 \simeq \sigma^2, \text{ for large } n. \qquad \ldots(*)$$

Also $$\text{Var}\left(\frac{ns^2}{\sigma^2}\right) = 2(n-1)$$

$$\Rightarrow \quad \frac{n^2}{\sigma^4} \text{Var}(s^2) = 2(n-1)$$

$$\Rightarrow \quad \text{Var}(s^2) = \frac{2}{n}\left(1 - \frac{1}{n}\right)\sigma^4 \simeq \frac{2\sigma^4}{n}, \text{ for large } n. \qquad \ldots(**)$$

$$\Rightarrow \quad \text{S.E.}(s^2) = \sigma^2 \times \sqrt{2/n} \qquad \ldots(***)$$

Theorem 13·6. *Let X_i, $(i = 1, 2, \ldots, n)$ be independent $N(0, 1)$ variates. Then the conditional distribution of $\chi^2 = \sum_{i=1}^{n} X_i^2$, subject to m $(< n)$ (say), independent homogeneous linear constraints viz.,*

$$\left.\begin{array}{l} c_{11}X_1 + c_{12}X_2 + \ldots\ldots + c_{1n}X_n = 0 \\ c_{21}X_1 + c_{22}X_2 + \ldots\ldots + c_{2n}X_n = 0 \\ \quad\vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \\ c_{m1}X_1 + c_{m2}X_2 + \ldots\ldots + c_{mn}X_n = 0 \end{array}\right\} \qquad \ldots(13\cdot12)$$

is also a χ^2-distribution with $(n - m)$ degrees of freedom.

Proof. Equivalently, the constraints (13·12) can be expressed as

$$\left.\begin{array}{l} a_{11}X_1 + a_{12}X_2 + \ldots\ldots + a_{1n}X_n = 0 \\ a_{21}X_1 + a_{22}X_2 + \ldots\ldots + a_{2n}X_n = 0 \\ \quad\vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \\ a_{m1}X_1 + a_{m2}X_2 + \ldots\ldots + a_{mn}X_n = 0 \end{array}\right\} \qquad \ldots(13\cdot12a)$$

where $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{in})$; $i = 1, 2, \ldots, m$ are m unitary, mutually orthogonal vectors.

Let us now transform the variables

$(X_1, X_2, \ldots, X_m, X_{m+1}, \ldots, X_n)$ to $(Y_1, Y_2, \ldots, Y_m, Y_{m+1}, \ldots, Y_n)$

by means of a linear orthogonal transformation

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \qquad \ldots(13\cdot12b)$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \\ Y_{m+1} \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \\ a_{m+1,1} & a_{m+1,2} & \cdots & a_{m+1,n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \\ X_{m+1} \\ \vdots \\ X_n \end{pmatrix}$$

(13·12b) implies that the constraints (13·12a) are equivalent to

$$Y_i = 0, \ (i = 1, 2, \ldots, m) \qquad \ldots(13\cdot12c)$$

By Fisher's Lemma (Theorem 13·4) Y_i, $(i = 1, 2, \ldots, n)$ are also independent $N(0, 1)$ variables and

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2 \qquad [\because \text{Transformation } (13\cdot2b) \text{ is orthogonal}]$$

$$= \sum_{i=m+1}^{n} Y_i^2 \qquad [\text{Using } (13\cdot12c)]$$

Thus the conditional distribution of $\sum_{i=1}^{n} X_i^2$ subject to the conditions (13·12) is same as the unconditional distribution of $\sum_{i=m+1}^{n} Y_i^2$; where Y_i $(i = m+1, m+2, \ldots, n)$ are independent standard normal variates without any constraints on them. Hence

$$\chi^2 = \sum_{i=1}^{n} X_i^2 = \sum_{i=m+1}^{n} Y_i^2,$$

being the sum of squares of $(n-m)$ independent standard normal variates follows χ^2-distribution with $(n-m)$ degrees of freedom.

Example 13·1. (a) *Show that for 2 d.f. the probability P of a value of χ^2 greater than χ_0^2 is $\exp(-\frac{1}{2}\chi_0^2)$, and hence that*

$$\chi_0^2 = 2 \log_e (1/P)$$

Deduce the value of χ_0^2 when $P = 0{\cdot}05$.

[*Sardar Patel Univ. B.Sc., 1991*]

(b) *Given different probabilities $P_1, P_2, \ldots, P_n$ obtained from n independent tests of significance, explain how you will pool them to get a single probability in order to decide about the significance of the aggregate of these tests.*

[*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

Solution. (*a*) The p.d.f. of χ^2-distribution with 2 d.f. is

$$f(\chi^2) = \left[\frac{1}{2^{n/2}\,\Gamma(n/2)} \exp(-\chi^2/2).(\chi^2)^{(n/2)-1}\right]_{n=2}$$

$$= \frac{1}{2}\exp(-\chi^2/2),\ 0 \le \chi^2 < \infty$$

$$\therefore \quad P = P(\chi^2 > \chi_0^2) = \int_{\chi_0^2}^{\infty} \frac{1}{2}\exp(-\chi^2/2)\,d\chi^2 \qquad \ldots(*)$$

$$= \frac{1}{2}\left|\frac{\exp(-\chi^2/2)}{-\frac{1}{2}}\right|_{\chi_0^2}^{\infty} = \exp(-\chi_0^2/2)$$

$$\therefore \quad \log_e P = -\chi_0^2/2$$

$$\Rightarrow \quad \chi_0^2 = -2\log_e P = 2\log_e(1/P)$$

When $P = 0{\cdot}05$, we get $\chi_0^2 = 2\log_e 20 \doteq 3{\cdot}012$

Remark. The value χ_0^2 of χ^2 defined in (*), is known as the significant or critical value [*c.f.* Remark 4, to Theorem 13·3, page 13·15] of χ^2 corresponding to the probability level P. Thus if P is the significant probability, then

$$\chi^2 = -2\log_e P = 2\log_e(1/P) \qquad \ldots(13{\cdot}13)$$

is a χ^2-variate with 2 d.f.

(*b*) $-2\log_e P_i\ (i = 1, 2, \ldots, n)$ are *independent* χ^2-variates each with 2 d.f. (*c.f.* Remark above and the fact that P_i's obtained from independent tests of significance are independent). Hence by additive property of chi-square distribution

$$\chi^2 = \sum_{i=1}^{n}(-2\log_e P_i) = 2\log_e\left(\frac{1}{P_1 P_2 \ldots P_n}\right) \qquad \ldots(13{\cdot}13a)$$

is a chi-square variate with $2n$ d.f.

If $\chi^2 > \chi^2_{0{\cdot}05}$ for $2n$ d.f., then we conclude that the pooled result (aggregate of the tests) is significant at 5% level of significance.

Example 13·2. (*Pearson's P_χ-Statistic*). *The variables $X_1, X_2, \ldots, X_n$ are independently distributed in the rectangular form*

$$dF = dx,\ 0 \le x \le 1$$

Then if $P = X_1 X_2 \ldots X_n$, show that $-2\log_e P$ has χ^2-distribution with $2n$ degrees of freedom. (*Aligarh Univ. B.Sc., 1992*)

Solution. $-2\log_e P = -2\log_e(X_1 X_2, \ldots X_n)$

$$= \xi_1 + \xi_2 + \ldots + \xi_n = \sum_{i=1}^{n} \xi_i ,$$

where $\xi_i = -2 \log X_i \Rightarrow X_i = \exp(-\xi_i/2)$.

The probability function of ξ_i is given by

$$g(\xi_i) = f(x_i) \left| \frac{dx_i}{d\xi_i} \right|$$

Since $\quad dF(x) = dx,\ f(x) = 1 \ \forall\ x$ in $[0, 1]$

$$\therefore \quad g(\xi_i) = 1 . \left| \exp(-\xi_i/2) \times \left(-\frac{1}{2}\right) \right| = \frac{1}{2} \exp(-\xi_i/2)$$

which is the probability function of χ^2-distribution with 2 d.f.

$\therefore$ ξ_i, $(i = 1, 2, \ldots, k)$ are independent χ^2-variates each with 2 d.f. Hence by additive property of χ^2-distribution,

$$-2 \log_e P = \sum_{i=1}^{n} \xi_i ,$$

is a χ^2-variate with $2n$ d.f.

Remark. The significance of this result lies in testing of hypothesis as explained in Example 13·1.

Example 13·3. *Show that if v is even,*

$$P = \frac{1}{2^{(v-2)/2}\Gamma(v/2)} \int_{\chi}^{\infty} \exp(-\chi^2/2)\, \chi^{v-1} d\chi$$

$$= \exp(-\chi^2/2) \left[1 + (\chi^2/2) + \frac{\chi^4}{2 \cdot 4} + \ldots + \frac{\chi^{v-2}}{2 . 4 \ldots (v-2)} \right]$$

and hence the values of P for a given χ^2 can be derived from tables of Poisson's exponential limit.

Solution. Let us consider the incomplete Gamma integral

$$I_r = \frac{1}{r\,!} \int_{\beta}^{\infty} e^{-t}\, t^r\, dt,$$

where r is a positive integer. Integrating by parts, we get

$$I_r = \left| -\frac{e^{-t} . t^r}{r\,!} \right|_{\beta}^{\infty} + \frac{1}{(r-1)\,!} \int_{\beta}^{\infty} e^{-t}\, t^{r-1}\, dt = \frac{e^{-\beta} . \beta^r}{r\,!} + I_{r-1}$$

which is a reduction formula. Repeated application of this gives

$$I_r = \frac{e^{-\beta} \beta^r}{r\,!} + \frac{e^{-\beta} \beta^{r-1}}{(r-1)\,!} + \ldots + \frac{e^{-\beta} . \beta^2}{2\,!} + \frac{e^{-\beta} . \beta}{1\,!} + I_0$$

But $\quad I_0 = \int_{\beta}^{\infty} e^{-t}\, dt = \left| -e^{-t} \right|_{\beta}^{\infty} = e^{-\beta}$

$$\therefore \quad \frac{1}{r\,!}\int_{\beta}^{\infty} e^{-t}t^{r}\,dt = e^{-\beta}\left[1+\beta+\frac{\beta^2}{2\,!}+\ldots+\frac{\beta^{r-1}}{(r-1)\,!}+\frac{\beta^r}{r\,!}\right]$$

Putting $\beta = \chi^2/2$ and $r = \dfrac{\nu-2}{2} = \dfrac{\nu}{2} - 1$, (since r is an integer, $\nu = 2r + 2$ must be even), we get

$$\frac{1}{\{(\nu/2)-1\}\,!}\int_{\chi^2/2}^{\infty} e^{-t}\,t^{(\nu/2)-1}dt$$

$$= \exp(-\chi^2/2)\left[1+\frac{\chi^2}{2}+\frac{\chi^4}{2.4}+\frac{\chi^6}{2.4.6}+\ldots+\frac{\chi^{\nu-2}}{2.4.6\ldots(\nu-2)}\right] \qquad \ldots(*)$$

Taking $t = \chi^2/2$ in the integral on the L.H.S., we get

$$\text{L.H.S.} = \frac{1}{\{(\nu/2)-1\}\,!}\int_{\chi}^{\infty} \exp(-\chi^2/2)\,(\chi^2/2)^{(\nu/2)-1}\,d(\chi^2/2)$$

$$= \frac{1}{2^{(\nu-2)/2}\,\Gamma(\nu/2)}\int_{\chi}^{\infty} \exp(-\chi^2/2)\,\chi^{\nu-1}\,d\chi \qquad \ldots(**)$$

From (*) and (**), we get the required result.

Let the given value of χ^2 be χ_0^2, then

$$P = P(\chi^2 > \chi_0^2) = \frac{1}{2^{(\nu-2)/2}\,\Gamma(\nu/2)}\int_{\chi_0}^{\infty} \exp(-\chi^2/2)\,\chi^{\nu-1}\,d\chi$$

$$= \exp(-\chi_0^2/2)\left[1+\frac{\chi_0^2}{2}+\frac{\chi_0^4}{2.4}+\ldots+\frac{\chi_0^{\nu-2}}{2.4\ldots(\nu-2)}\right]$$

$$= e^{-\lambda}\left[1+\lambda+\frac{\lambda^2}{2\,!}+\frac{\lambda^3}{3\,!}+\ldots+\frac{\lambda^{\frac{\nu}{2}-1}}{[(\nu/2)-1]\,!}\right],$$

where $\lambda = \chi_0^2/2$.

The terms on the right hand side *viz.*, $e^{-\lambda}, \lambda e^{-\lambda}, \dfrac{\lambda^2}{2\,!}e^{-\lambda}\ldots$etc. are the successive terms of the Poisson distribution with parameter $\lambda = \chi_0^2/2$.

Hence the result.

Example 13·4. *If X and Y are independent normal variates with means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively, derive the distribution of*

$$Z = (X - \mu_1)/(Y - \mu_2).$$

What is the name of the distribution so obtained ? Mention one important property of this distribution.

Solution. Here $Z^2 = \dfrac{(X-\mu_1)^2}{(Y-\mu_2)^2} \Rightarrow \dfrac{\sigma_2^2}{\sigma_1^2}\cdot Z^2 = \dfrac{[(X-\mu_1)/\sigma_1]^2}{[(Y-\mu_2)/\sigma_2]^2}$

But $\{(X-\mu_1)/\sigma_1\}^2$ and $\{(Y-\mu_2)/\sigma_2\}^2$, being the squares of independent standard normal variates, are independent χ^2-variates with 1 d.f. each.

Thus $\frac{\sigma_2^2 Z^2}{\sigma_1^2}$, being the quotient of two independent χ^2-variates each with 1 d.f. is a $\beta_2\left(\frac{1}{2}, \frac{1}{2}\right)$ variate (c.f. *Theorem 13·1*).

Hence its probability differential is given by

$$dF\left(\frac{\sigma_2^2}{\sigma_1^2} z^2\right) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \times \frac{(\sigma_2^2 z^2/\sigma_1^2)^{(1/2)-1}}{\left(1 + \frac{\sigma_2^2 z^2}{\sigma_1^2}\right)^{\frac{1}{2}+\frac{1}{2}}} d(\sigma_2^2 z^2/\sigma_1^2)$$

$$= \frac{\Gamma\left(\frac{1}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \cdot \frac{\left(\frac{\sigma_2^2 z^2}{\sigma_1^2}\right)^{\frac{1}{2}-1}}{\left(1 + \frac{\sigma_2^2 z^2}{\sigma_1^2}\right)} \cdot \frac{\sigma_2^2}{\sigma_1^2} d(z^2)$$

$$= \frac{\sigma_1 \sigma_2}{\pi(\sigma_1^2 + \sigma_2^2 z^2) z} dz^2, \quad 0 < z^2 < \infty \qquad [\because \Gamma(1/2) = \sqrt{\pi}]$$

Thus the probability differential of Z is given by

$$dF(z) = f(z)dz = \frac{\sigma_1 \sigma_2}{\pi(\sigma_1^2 + \sigma_2^2 z^2)} dz, \quad -\infty < z < \infty$$

If $\sigma_1 = \sigma_2 = 1$, then it conforms to standard Cauchy distribution,

$$dF(z) = \frac{1}{\pi} \cdot \frac{dz}{(1 + z^2)}, \quad -\infty < z < \infty$$

[For its properties *c.f.* Chapter 8].

Aliter $\qquad Z = \frac{X - \mu_1}{Y - \mu_2}$,

$$\Rightarrow \qquad \frac{\sigma_2}{\sigma_1} Z = \frac{(X - \mu_1)/\sigma_1}{(Y - \mu_2)/\sigma_2}$$

Now $\frac{\sigma_2}{\sigma_1} Z$, being the ratio of two independent standard normal variates is a standard Cauchy variate

$$\therefore \qquad dF\left(\frac{\sigma_2}{\sigma_1} z\right) = \frac{d\left(\frac{\sigma_2}{\sigma_1} z\right)}{\pi\left[1 + \left(\frac{\sigma_2}{\sigma_1} z\right)^2\right]}$$

$$= \frac{\sigma_1 \sigma_2}{\pi(\sigma_1^2 + \sigma_2^2 z^2)} dz, \quad -\infty < z < \infty$$

Example 13·5. *X_i, $(i = 1, 2, \ldots, n)$ are independently and normally distributed with zero mean and common variance σ^2.*

Let $\xi_i = \sum_{j=1}^{n} c_{ij} X_j$; $i = 1, 2, \ldots, n$, *where* $\sum_{j=1}^{n} c_{ij} c_{i'j} = \delta_{ii'}$

where $\delta_{ii'}$ is Kroneker delta. Show that

$$\left[\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{p} \xi_i^2\right] \Big/ \sigma^2$$

is distributed as χ^2-variate with $(n-p)$ degrees of freedom.

[*Delhi Univ. M.Sc. (Stat.),* 1990]

Solution. Since $\delta_{ii'} = \sum_{j=1}^{n} c_{ij} c_{i'j}$

is a Kroneker delta, we have

$$\sum_{j=1}^{n} c_{ij} c_{i'j} = \begin{cases} 0, & i \neq i' \\ 1, & i = i' \end{cases}$$

i.e., X_i's are transformed to ξ_i's by means of a linear orthogonal transformation. Hence by Fisher's Lemma, ξ_i, $(i = 1, 2, \ldots, n)$ are independent normal variates with zero mean and common variance σ^2.

Since the transformation is orthogonal, we have

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} \xi_i^2,$$

$$\therefore \quad \frac{\left[\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{p} \xi_i^2\right]}{\sigma^2} = \frac{\sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{p} \xi_i^2}{\sigma^2} = \frac{\sum_{i=p+1}^{n} \xi_i^2}{\sigma^2}$$

Now $$\frac{\sum_{i=p+1}^{n} \xi_i^2}{\sigma^2} = \sum_{i=p+1}^{n} (\xi_i/\sigma)^2,$$

being the sum of the squares of $(n-p)$ independent standard normal variates is a χ^2-variate with $(n-p)$ degrees of freedom. Hence the result.

Example 13·6. *Show that the m.g.f. of $Y = \log \chi^2$, where χ^2 follows chi-square distribution with n d.f., is given by*

$$M_Y(t) = 2^t\, \Gamma\left(\frac{n}{2} + t\right) \Big/ \Gamma(n/2)$$

If χ_1^2 and χ_2^2 are independent χ^2-variates each with n.d.f. and $U = \chi_1^2/\chi_2^2$, deduce that for positive integer k,

$$E(U^k) = \Gamma\left(\frac{n}{2} + k\right)\Gamma\left(\frac{n}{2} - k\right) \Big/ \left[\Gamma\left(\frac{n}{2}\right)\right]^2$$

Solution. $y = \log \chi^2 \Rightarrow \chi^2 = e^y \Rightarrow d\chi^2 = e^y\, dy$.

The probability differential of χ^2 *viz.,*

$$dP(\chi^2) = \frac{1}{2^{n/2}\,\Gamma(n/2)}\, e^{-\chi^2/2} (\chi^2)^{\frac{n}{2}-1}\, d\chi^2,\ 0 < \chi^2 < \infty$$

transforms to

$$dG(y) = \frac{1}{2^{n/2}\,\Gamma(n/2)} \exp\left[-\frac{1}{2}e^y + \frac{ny}{2}\right] dy, \ -\infty < y < \infty$$

$$M_Y(t) = \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}e^y + \frac{ny}{2} + ty\right] dy$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^{\infty} e^{-z}\,(2z)^{\frac{n}{2}+t}\,\frac{dz}{z}, \qquad (2z = e^y)$$

$$= \frac{2^t}{\Gamma(n/2)} \int_0^{\infty} e^{-z}\, z^{\frac{n}{2}+t-1}\, dz$$

$$= 2^t\,\Gamma\left(\frac{n}{2}+t\right) \Big/ \Gamma\left(\frac{n}{2}\right) \qquad \ldots(*)$$

$$E(U^k) = E\left[\left(\frac{\chi_1^2}{\chi_2^2}\right)^k\right] = E\left[\exp\left\{\log\left(\frac{\chi_1^2}{\chi_2^2}\right)^k\right\}\right]$$

$$= E\left[e^{k \log \chi_1^2 - k \log \chi_2^2}\right]$$

$$= E\left(e^{k \log \chi_1^2}\right) . E\left(e^{-k \log \chi_2^2}\right)$$

[$\because \chi_1^2$ and χ_2^2 are independent]

$$= M_{\log \chi_1^2}(k) \,.\, M_{\log \chi_2^2}(-k)$$

$$= \frac{2^k\,\Gamma\left(\frac{n}{2}+k\right)}{\Gamma(n/2)} \cdot \frac{2^{-k}\,\Gamma\left(\frac{n}{2}-k\right)}{\Gamma(n/2)} \qquad \text{[From (*)]}$$

$$= \Gamma\left(\frac{n}{2}+k\right)\Gamma\left(\frac{n}{2}-k\right) \Big/ \left[\Gamma\left(\frac{n}{2}\right)\right]^2$$

Example 13·7. *If X is chi-square variate with n.d.f., then prove that for large n,* $\sqrt{2X} \sim N(\sqrt{2n}\,,\, 1)$ **[Delhi Univ. B.Sc. (Stat. Hons.), 1989]**

Solution. We have $E(X) = n$, $\text{Var}(X) = 2n$

$$Z = \frac{X - E(X)}{\sigma_X} = \frac{X - n}{\sqrt{2n}} \sim N(0, 1), \text{ for large } n.$$

Consider,

$$P\left(\frac{X-n}{\sqrt{2n}} \le z\right) = P(X \le n + z\sqrt{2n})$$

$$= P\,[\sqrt{2X} \le (2n + 2z\sqrt{2n})^{1/2}]$$

$$= P\left[\sqrt{2X} \le \sqrt{2n}\left(1 + z\sqrt{\frac{2}{n}}\right)^{1/2}\right]$$

$$= P\left[\sqrt{2X} \le \sqrt{2n}\left(1 + \frac{z}{\sqrt{2n}} - \frac{z^2}{4n} + \ldots\right)\right]$$

$$\approx P\,[\sqrt{2X} \le \sqrt{2n} + z], \text{ for large } n.$$
$$= P[\sqrt{2X} - \sqrt{2n} \le z], \text{ for large } n.$$

Since for large n, $(X - n)/\sqrt{2n} \sim N(0, 1)$, we conclude that

$$\sqrt{2X} - \sqrt{2n} \sim N(0, 1) \text{ for large } n.$$

$\Rightarrow$ $\sqrt{2X}$ is asymptotically $N(\sqrt{2n},\ 1)$.

Remark. This approximation is often used for the value of n larger than 30. This result does not reflect anything as to how good the approximation is, for moderate values of n. R.A. Fisher has proved that the approximation is improved by taking $\sqrt{(2n-1)}$ instead of $\sqrt{2n}$. A still better approximation is

$$(\chi^2/n)^{1/3} \sim N\left(1 - \frac{2}{9n}, \frac{2}{9n}\right)$$

Example 13·8. *For a chi-square distribution with n d.f. establish the following recurrence relation between the moments :*

$$\mu_{r+1} = 2r(\mu_r + n\mu_{r-1}),\ r \ge 1.$$

Hence find β_1 *and* β_2.

[*Delhi Univ.B.Sc. (Stat. Hons.), 1991*]

Solution. If $X \sim \chi^2_{(n)}$ then its *m.g.f.* about origin is

$$M_X(t) = E(e^{tX}) = (1-2t)^{-n/2}\ ;\ t < \frac{1}{2} \qquad \ldots(*)$$

Also $\quad E(X) = n = \mu$ (say).

Hence m.g.f. about mean, say, $M(t)$ is

$$M(t) = M_{X-\mu}(t) = E(e^{t(X-\mu)}) = e^{-\mu t}\,.\,E(e^{tX})$$
$$= e^{-nt}(1-2t)^{-n/2}$$

Taking logarithms of both sides, we get

$$\log M(t) = -nt - \frac{n}{2}\log(1-2t) \qquad \text{[Using (*)]}$$

Differentiating *w.r.* to t, we have

$$\frac{M'(t)}{M(t)} = -n + \frac{n}{2}\cdot\frac{2}{(1-2t)} = \frac{2nt}{(1-2t)}$$

$$\Rightarrow \qquad (1-2t)\,M'(t) = 2nt\,M(t)$$

Differentiating r times *w.r.* to t by Leibnitz theorem, we get

$$(1-2t)\,M^{r+1}(t) + r(-2)\,M^r(t) = 2nt\,M^r(t) + 2nr\,M^{r-1}(t)$$

Putting $t = 0$ and using the relation,

$$\mu_r = \left[\frac{d^r}{dt^r}\ M(t)\right]_{t=0} = M^r(0), \text{ we get}$$

$$\mu_{r+1} - 2r\,\mu_r = 2nr\,\mu_{r-1}$$

$$\Rightarrow \qquad \mu_{r+1} = 2r\,(\mu_r + n\mu_{r-1}),\ r \ge 1.$$

Substituting $r = 1, 2, 3$; we get

$$\mu_2 = 2n\mu_0 = 2n$$

$$\mu_3 = 4(\mu_2 + n\mu_1) = 8n \qquad [\because \mu_1 = 0 \text{ and } \mu_0 = 1]$$

$$\mu_4 = 6(\mu_3 + n\mu_2) = 48n + 12n^2$$

$$\therefore \qquad \beta_1 = \frac{\mu_3^2}{\mu_2^2} = \frac{8}{n} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{12}{n}$$

EXERCISE 13(a)

1. (*a*) Derive the p.d.f. of chi-square distribution with n degrees of freedom.

(*b*) If X has a chi-square distribution with n d.f., find m.g.f. $M_X(t)$.

Deduce that :

(*i*) $\mu_r' = EX^r = 2^r\,\Gamma[(n/2) + r] \,/\, \Gamma(n/2)$

(*ii*) $k_r = r$th cumulant $= n\, 2^{r-1}\,(r-1)\,!$

(*iii*) $k_1\, k_3 = 2k_2^2,\ 2\beta_2 - 3\beta_1 - 6 = 0.$

2. If $X \sim \chi^2_{(n)}$, show that :

(*i*) Mode is at $x = n - 2$.

(*ii*) The points of inflexion are equidistant from the mode.

Hint. Points of inflexion are at $x = (n-2) \pm [2(n-2)]^{1/2}$

3. If $X \sim \chi^2_{(n)}$, obtain the m.g.f. of X. Hence find the m.g.f. of standard chi-square variate and obtain its limiting form as $n \to \infty$. Also interpret the result.

4. (*a*) Let $X \sim N(0, 1)$ and $Y = X^2$. Calculate $E(Y)$ in two different ways.

Ans. $E(Y) = 1$. (Use Normal distribution and chi-square distribution).

(*b*) Let X_1 and X_2 be independent standard normal variates and let

$Y = (X_2 - X_1)^2/2$. Find the distribution of Y.

Ans. $Y \sim \chi^2_{(1)}$.

5. If $X_1, X_2, \ldots, X_n$ are *i.i.d.* exponential variates with parameter λ, prove that

$$2\lambda \sum_{i=1}^{n} X_i \sim \chi^2_{(2n)}$$

6. (*a*) If $X \sim$ U [0 1], show that $-2 \log X \sim \chi^2_{(2)}$.

Hence show that if $X_1, X_2, \ldots, X_n$ are *i.i.d* U [0, 1] variates, and if $P = X_1\, X_2 \ldots X_n$, then $-2 \log_e P \sim \chi^2_{(2n)}$

Hint. Find m.g.f. of $-2 \log X$.

(*b*) If $X_1, X_2, \ldots, X_n$ are independent random variables with continuous distribution functions $F_1, F_2, \ldots, F_n$ respectively, show that

$$-2 \log [F_1(X_1) \,.\, F_2(X_2) \ldots F_n(X_n)] \sim \chi^2_{(2n)}$$

Hint. Use $F(X) \sim$ U [0, 1] and Part (*a*) above.

7. (*a*) Let X and Y be two independent random variables having chi-square distribution with degrees of freedom m and n respectively.

(*i*) Obtain the distribution of $U = \dfrac{X}{X+Y}$,

(*ii*) When $m = n$, show that the distribution of U is symmetrical about $\frac{1}{2}$. Hence or otherwise derive the rth moment about the mean of U when $m = n$.

(*iii*) Deduce the distribution of U when $m = n = 1$.

(*b*) If X and Y are independently distributed chi-square variates with m and n degrees of freedom respectively, show that $U = X + Y$ and $V = X/Y$ are independently distributed. [*Gujarat Univ. B.Sc., 1992*]

(*c*) If χ_1^2 and χ_2^2 are independent χ^2 variates with n_1 and n_2 degrees of freedom respectively, then show that:

(*i*) $\chi^2 = \chi_1^2 + \chi_2^2$ is a χ^2-variate with $(n_1 + n_2)$ degrees of freedom

(*ii*) $T^2 = \dfrac{\chi_1^2}{\chi_2^2}$ is a $\beta_2\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)$ variate.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1987*]

8. If X_i $(i = 1, 2, \ldots, n)$ are n independent normal variates with zero means and unit variances, show that $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} (X_i - \bar{X})^2$ are independently distributed.

Hence or otherwise obtain the distribution of

$$U = \frac{\sum_{i=1}^{n} X_i}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

9. (*a*) Prove that $\dfrac{nS^2}{\sigma^2}$ is distributed like χ^2 with $(n-1)$ degrees of freedom, where S^2 and σ^2 are the variances of sample (of size n) and the population respectively. [*Burdwan Univ. B.Sc. (Maths.) Hons.), 1992*]

(*b*) Let X/σ_1^2 and Y/σ_2^2 be two independent chi-square variates with n and m degrees of freedom respectively. Find an unbiased estimate of $(\sigma_1/\sigma_2)^2$ and find its variance. Show that X/Y and $(X/\sigma_1^2) + (Y/\sigma_2^2)$ are independently distributed. Name the distributions of X/Y and $(X/\sigma_1^2) + (Y/\sigma_2^2)$.

10. X denotes the random variable with chi-square distribution having n degrees of freedom. Show that for suitably chosen constants a_n and b_n, the moment generating function of $\dfrac{X - a_n}{b_n}$ tends to that of the standard normal distribution as $n \to \infty$. From this what would you conclude about the behaviour, for large n, of $P\left(\dfrac{X - a_n}{b_n} \leq x\right)$?

11. Show that

$$P\{\chi^2_{2\nu+2} \geq 2\lambda\} = \frac{1}{\nu\,!}\int_{\lambda}^{\infty} e^{-y}\, y^{\nu} dy = \sum_{r=0}^{\nu} \frac{e^{-\lambda}\lambda^r}{r\,!}$$

where $\lambda = \frac{1}{2}\chi_0^2$.

Explain the uses of this result.[*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

12. X is a Poisson variate with parameter λ and χ^2 is a chi-square variate with $2k$ d.f. Prove that for all positive integers k,

$$P\{X \leq k-1\} = P\{\chi^2 > 2\lambda\}$$

Hint. $P(\chi^2 > 2\lambda) = \dfrac{1}{2^k\,\Gamma(k)} \displaystyle\int_{2\lambda}^{\infty} \exp\left(-\tfrac{1}{2}\chi^2\right)(\chi^2)^{k-1}\, d\chi^2$

$$= \frac{1}{(k-1)\,!}\int_{\lambda}^{\infty} e^{-y}\, y^{k-1}\, dy$$

$$= \frac{1}{(k-1)\,!}\left\{\lambda^{k-1} e^{-\lambda} + (k-1)\int_{\lambda}^{\infty} e^{-y}\, y^{k-2}\, dy\right\}$$

By repeated integration, we get the required result.

13. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent random samples from a normal population with mean zero and variance σ^2. Let their means be $\bar{X}$ and $\bar{Y}$ and their variances be S_X^2 and S_Y^2 respectively.

Let the pooled variance S_p^2 be defined as :

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)\,S_Y^2}{(m+n-2)}$$

Prove that $(\bar{X} - \bar{Y})$ and $(m+n-2)\,S_p^2/\sigma^2$ are independently distributed, the former as a normal variate with zero mean and variance $\sigma^2\,\{(1/m) + (1/n)\}$ and the latter as a chi-square variate with $(m+n-2)$ d.f.

[*Nagpur Univ B.E., 1992*]

14. If X is a random variable following Poisson distribution with parameter λ, and λ is also a random variable so that $2\alpha\lambda$ is a chi-square variate with $2p$ degrees of freedom, obtain the unconditional distribution of X. Give the name of this distribution and find its mean.

15. X_1, X_2, and X_3 denote independent central chi-square variates with ν_1, ν_2 and ν_3 d.f. respectively.

(*i*) Show that $X_1/(X_1 + X_2)$ is independently distributed of

$$(X_1 + X_2)/(X_1 + X_2 + X_3).$$

(*ii*) Obtain the joint density function of the distribution of

$$X = X_1/(X_1 + X_2 + X_3) \;\text{ and }\; Y = X_2/(X_1 + X_2 + X_3)$$

(*iii*) Hence or otherwise obtain the mean and variance of X and Y and Cov (X, Y).

16. Prove that each linear constraint on $\{f_i\}$, $i = 1, 2, \ldots, n$ reduces by unity the number of degrees of freedom of the chi-square,

$$\sum_{i=1}^{n} \{(f_i - e_i)^2/e_i\},$$

where $e_i = E(f_i)$.

17. If X follows a chi-square distribution with n d.f. so that $E(X) = n$ and $V(X) = 2n$, prove that $(X - n)/\sqrt{2n}$ is a $N(0, 1)$, for large n.

18. If $y\,dx$ is the probability that X lies between x and $x + dx$ and if y is given by the solution of the differential equation

$$\frac{dy}{dx} = \frac{y(a-x)}{bx+c}$$

show that, (for suitable values of the constants a, b and c), a certain linear function of X has the χ^2-distribution with n degrees of freedom, where

$$n = 2\left(1 + \frac{a}{b} + \frac{c}{b^2}\right)$$

19. If X_1 and X_2 are independently distributed, each as χ^2 variate with 2 d.f., show that the density function of $Y = \frac{1}{2}(X_1 - X_2)$ is

$$g(y) = \frac{1}{2}\, e^{-|y|},\ -\infty < y < \infty.$$

20. If X and Y are independent r.v.'s having rectangular distribution in the interval $(0, 1)$, show that

$$U = \sqrt{-2 \log X} \cos 2\pi Y \quad \text{and} \quad V = \sqrt{-2 \log X} \sin 2\pi Y$$

are independently distributed as $N(0, 1)$. Hence show that U^2 and V^2 are independently distributed as χ^2-variates, each with 1 d.f.

Hint. $\dfrac{1}{J} = \dfrac{\partial(u,v)}{\partial(x,y)} = -\dfrac{2\pi}{x}$

and $u^2 + v^2 = -2 \log x \;\Rightarrow\; x = \exp\left[-\frac{1}{2}(u^2 + v^2)\right]$

21. Find the p.d.f. of $\chi_n = +\sqrt{\chi_n^2}$, where χ_n^2 is a χ^2-variate with n d.f. and show that

$$\mu_r' = E(\chi_n^r) = 2^{r/2}\,\frac{\Gamma[(n+r)/2]}{\Gamma(n/2)}$$

Hence establish that for large n,

$$E(\chi_n^2) \simeq [E(\chi_n)]^2$$

[**Hint.** $E(\chi_n^2) = n$ and $E(\chi_n) = \mu_1' = 2^{1/2}\,\dfrac{\Gamma[(n+1)/2]}{\Gamma(n/2)}$

Now use $\dfrac{\Gamma(n+k)}{\Gamma n} \simeq n^k$, for large values of n. [c.f. Remark to § 14·5·7]

22. Let

$$P_x = \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^x w^{(n-2)/2} \cdot e^{-w/2}\, dw,\ x > 0.$$

Show that

$$x < \frac{n}{1 - P_x}$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

23. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, and k be a positive integer. Find $E(S^{2k})$. In particular, find $E(S^2)$ and Var (S^2).

$$\left[S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \; ; \; \bar{X} = n^{-1} \sum_{i=1}^{n} X_i\right]$$

Ans. $E(S^{2k}) = \left(\dfrac{2\sigma^2}{n-1}\right)^k \cdot \dfrac{\Gamma\left(k + \dfrac{n-1}{2}\right)}{\Gamma\left(\dfrac{n-1}{2}\right)}$; $k > 0, n > 1$

$E(S^2) = \sigma^2$, Var $(S^2) = 2\sigma^4/(n-1)$.

24. Let X_1 and X_2 be independent random variables, each $N(0, 1)$. Find the joint distribution of $Y_1 = X_1^2 + X_2^2$ and $Y_2 = X_1/X_2$. Find the marginal distributions of Y_1 and Y_2. Are Y_1 and Y_2 independent ?

Ans. $Y_1 \sim \chi^2_{(2)}$ and Y_2 is standard Cauchy variate. Yes.

25. Let X_1 and X_2 be independent standard normal variates. Let

$$Y_1 = X_1 + X_2 \text{ and } Y_2 = X_1^2 + X_2^2.$$

(*i*) Show that the joint m.g.f. of Y_1 an Y_2 is :

$$M(t_1, t_2) = \frac{1}{1 - 2t_2} \exp\left[\frac{t_1^2}{1 - 2t_2}\right]; \; -\infty < t_1 < \infty, -\infty < t_2 < \frac{1}{2}$$

(*ii*) Hence or otherwise, show that
Y_1 is a normal variate and Y_2 is a chi-square variate.

(*iii*) Are Y_1 and Y_2 independent ? If not, find the correlation coefficient of Y_1 and Y_2.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1989*]

Ans. (Y_1, Y_2) are not independent. $\rho(Y_1, Y_2) = 0$.

26. If $X_1, X_2, \ldots, X_n$ are independently and normally distributed with the same mean but different variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ and assuming that

$$U = \frac{\sum_i (X_i/\sigma_i^2)}{\sum_i (1/\sigma_i^2)} \quad \text{and} \quad V = \sum_{i=1}^{n} \left[\frac{(X_i - U)^2}{\sigma_i^2}\right]$$

are independently distributed, show that $U \sim N\{0, 1/(\sum_i \sigma_i^2)\}$ and V has χ^2 distribution with $(n-1)$ d.f.

27. If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, find the mean and variance of

$$S = \left[\sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1) \right]^{1/2}$$

[*Delhi Univ. B.Sc. (Maths. Hons.), 1988*]

28. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(0, 1)$. Define :

$$\overline{X}_k = \frac{1}{k} \sum_{i=1}^{k} X_i \text{ and } \overline{X}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^{n} X_i$$

(*a*) What is the distribution of $\frac{1}{2}(\overline{X}_k + \overline{X}_{n-k})$?

(*b*) What is the distribution of $k\overline{X}_k^2 + (n-k)\overline{X}^2_{n-k}$?

(*c*) What is the distribution of X_i^2/X_j^2, $i \neq j$?

(*d*) What is the distribution of X_i/X_j, $i \neq j$?

Ans. (*a*) $N\left(0, \dfrac{n}{4k(n-k)}\right)$; (*b*) $\chi^2_{(2)}$

(*c*) $\beta_2\left(\frac{1}{2}, \frac{1}{2}\right)$ or $F(1, 1)$ [See § 14·5]

(*d*) Standard Cauchy distribution.

OBJECTIVE TYPE QUESTIONS

I. Choose the correct answer from *B* and match it with each item in *A*.

	A		*B*
(*a*)	β_2 for a chi-square distribution	(1)	$(1-2it)^{-n/2}$
(*b*)	β_1 for a chi-square distribution	(2)	$8/n$
(*c*)	Mean for a chi-square distribution	(3)	$2n$
(*d*)	Variance for a chi-square distribution	(4)	$(1-2t)^{-n/2}$
(*e*)	Characteristic function for χ^2 distribution	(5)	$(12/n) + 3$
(*f*)	Mode of χ^2-distribution	(6)	$\sqrt{2/n}$
(*g*)	M.G.F. of χ^2-distribution	(7)	n
(*h*)	Skewness of χ^2-distribution	(8)	$(n-2)$

II. State which of the following statement are True and which are False. In case of false statements, give the correct statement.

(*i*) Normal distribution is particular case of χ^2-distribution for one d.f.

(*ii*) For large degrees of freedom, chi-square distribution tends to normal distribution.

(*iii*) The sum of independent chi-square variates is also a chi-square variate.

(*iv*) For the validity of χ^2-test, it is always necessary that the sample observations should be independent.

(*v*) The chi-square distribution maintains its character of continuity if cell frequency is less than 5.

(*vi*) Each linear constraint reduces the number of degrees of freedom of chi-square by unity.

(*vii*) In a chi-square test of goodness of fit, if the calculated value of χ^2 is zero then the fit is a bad fit.

III. Mention the correct answer :

(*i*) The mean of a chi-square distribution with n d.f. is

(*a*) $2n$, (*b*) n^2, (*c*) $\sqrt{n}$, (*d*) n

(*ii*) The characteristic function of chi-square distribution is

(*a*) $(1-2\,it)^{n/2}$, (*b*) $(1+2\,it)^{n/2}$, (*c*) $(1-2\,it)^{-n/2}$

(*iii*) The range of χ^2-variate is

(*a*) $-\infty$ to $+\infty$, (*b*) 0 to ∞, (*c*) 0 to 1, (*d*) $-\infty$ to 0.

(*iv*) The skewness in a chi-square distribution will be zero if

(*a*) $n \to \infty$, (*b*) $n = 0$, (*c*) $n = 1$, (*d*) $n < 0$

(*v*) The moment generating function of a χ^2-distribution with n degrees of freedom is

(*a*) $(1-t)^{-n/2}$, (*b*) $(1-2t)^{-n/2}$, (*c*) $(1-3t)^{-n/2}$, (*d*) $(1-2t)^{n/2}$

(*iv*) Chi-square distribution is

(*a*) Continuous, (*b*) multimodal, (*c*) symmetrical.

IV. Mention some prominent features of the chi-square distribution with n degrees of freedom.

V. If X and Y are independent random variables having chi-square distribution with m and n degrees of freedom respectively, write down the distributions of (*i*) $X+Y$, (*ii*) X/Y, (*iii*) $X/(X+Y)$.

VI. (*a*) For how many degrees of freedom does the χ^2-distribution reduce to negative exponential distribution ?

(*b*) Give an example of two independent variates none of which is a chi-square variate, although their sum is a chi-square variate.

13·7. Applications of Chi-square Distribution. χ^2-distribution has a large number of applications in Statistics, some of which are enumerated below :

(*i*) To test if the hypothetical value of the population variance is $\sigma^2 = \sigma_0^2$ (say).

(*ii*) To test the 'goodness of fit'.

(*iii*) To test the independence of attributes.

(*iv*) To test the homogeneity of independent estimates of the population variance.

(*v*) To combine various probabilities obtained from independent experiments to give a single test of significance.

(*vi*) To test the homogeneity of independent estimates of the population correlation coefficient.

In the following sections we shall briefly discuss these applications.

13·7·1. Chi-square Test for Population Variance. Suppose we want to test if a random sample x_i, $(i = 1, 2, ..., n)$ has been drawn from a normal population with a specified variance $\sigma^2 = \sigma_0^2$, (say).

Under the null hypothesis that the population variance is $\sigma^2 = \sigma_0^2$, the statistic

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})^2}{\sigma_0^2}\right] = \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] = ns^2/\sigma_0^2 \quad ...(13·14)$$

follows chi-square distribution with $(n-1)$ d.f.

By comparing the calculated value with the tabulated value of χ^2 for $(n-1)$ d.f. at certain level of significance, (usually 5%), we may retain or reject the null hypothesis.

Remarks. 1. The above test (13·14) can be applied only if the population from which sample is drawn is normal.

2. If the sample size n is large (>30), then we can use Fisher's approximation

$$\sqrt{2\chi^2} \sim N(\sqrt{2n-1},\ 1)$$

i.e.,

$$Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0, 1) \quad ...(13·14a)$$

and apply Normal Test.

3. For a detailed discussion on the *significant values, (critical values)*, for testing $H_0 : \sigma^2 = \sigma_0^2$ against various alternatives : *(i)* $\sigma^2 > \sigma_0^2$; *(ii)* $\sigma^2 < \sigma_0^2$ and *(iii)* $\sigma^2 \neq \sigma_0^2$, see Remark 1 to § 16·7·4.

Example 13·9. *It is believed that the precision (as measured by the variance) of an instrument is no more than 0·16. Write down the null and alternative hypothesis for testing this belief. Carry out the test at 1% level, given 11 measurements of the same subject on the instrument :*

2·5, 2·3, 2·4, 2·3, 2·5, 2·7, 2·5, 2·6, 2·6, 2·7, 2·5.

[***Calicut Univ. B.Sc. (Main Stat.), April 1989***]

Solution. *Null Hypothesis,* $H_0 : \sigma^2 = 0·16$

Alternative Hypothesis : $H_1 : \sigma^2 > 0·16$

COMPUTATION OF SAMPLE VARIANCE

X	$X - \bar{X}$	$(X - \bar{X})^2$
2·5	− 0·01	0·0001
2·3	− 0·21	0·0441
2·4	− 0·11	0·0121
2·3	− 0·21	0·0441
2·5	− 0·01	0·0001
2·7	+ 0·19	0·0361
2·5	− 0·01	0·0001
2·6	+ 0·09	0·0081
2·6	+ 0·09	0·0081
2·7	+ 0·19	0·0361
2·5	− 0·01	0·0001
$\bar{X} = \frac{27·6}{11} = 2·51$		$\Sigma(X - \bar{X})^2 = 0·1891$

Under the null hypothesis $H_0 : \sigma^2 = 0 \cdot 16$, the test statistic is :

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\Sigma(X - \bar{X})^2}{\sigma^2} = \frac{0 \cdot 1891}{0 \cdot 16} = 1 \cdot 182$$

which follows χ^2-distribution with *d.f.* $(11 - 1) = 10$.

Since the calculated value of χ^2 is less than the tabulated value 23·2 of χ^2 for 10 *d.f.* at 1% level of significance, it is not significant. Hence H_0 may be accepted and we conclude that the data are consistent with the hypothesis that the precision of the instrument is 0·16.

Example 3·10. *Test the hypothesis that $\sigma = 10$, given that $s = 15$ for a random sample of size 50 from a normal population.*

Solution. *Null Hypothesis,* $H_0 : \sigma = 10$.

We are given $n = 50$, $s = 15$

$$\therefore \qquad \chi^2 = \frac{ns^2}{\sigma^2} = \frac{50 \times 225}{100} = 112 \cdot 5$$

Since n is large, using(13·14a), the test statistic is

$$Z = \sqrt{2\chi^2} - \sqrt{2n - 1} \sim N(0, 1)$$

Now, $Z = \sqrt{225} - \sqrt{99} = 15 - 9 \cdot 95 = 5 \cdot 05$

Sinc $|Z| > 3$, it is significant at all levels of significance and hence H_0 is rejected and we conclude that $\sigma \neq 10$.

13·7·2. Chi-square Test of Goodness of Fit. A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as "*Chi-square test of goodness of fit.*" It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If O_i, $(i = 1, 2, ..., n)$ is a set of observed (experimental) frequencies and E_i $(i = 1, 2, ..., n)$ is the corresponding set of expected (theoretical or hypothetical) frequencies, then Karl Pearson's chi-square, given by

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right], \qquad \left(\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i \right) \qquad ...(13 \cdot 15)$$

follows chi-square distribution with $(n - 1)$ d.f.

Remark. This is an approximate test for large values of n. The conditions for the validity of the χ^2-test of goodness of fit have already been given in § 13·5 on page 13·15.

Example 13·11. *The following figures show the distribution of digits in numbers chosen at random from a telephone directory :*

Digits :	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*
Frequency :	*1026*	*1107*	*997*	*966*	*1075*	*933*	*1107*	*972*	*964*	*853*	*10,000*

Test whether the digits may be taken to occur equally frequently in the directory. **[*Osmania Univ. M.A. (Eco.), 1992*]**

Solution. Here we set up the *null hypothesis* that the digits occur equally frequently in the directory.

Under the null hypothesis, the expected frequency for each of the digits 0, 1,2, ..., 9 is 10000/10 = 1000. The value of χ^2 is computed as follows :

CALCULATIONS FOR χ^2

Digits	*Observed Frequency (O)*	*Expected Frequency (E)*	$(O-E)^2$	$(O-E)^2/E$
0	1026	1000	676	0·676
1	1107	1000	11449	11·449
2	997	1000	9	0·009
3	966	1000	1156	1·156
4	1075	1000	5625	5·625
5	933	1000	4489	4·489
6	1107	1000	11449	11·449
7	972	1000	784	0·784
8	964	1000	1296	1·296
9	853	1000	21609	21·609
Total	10,000	10,000		58·542

$$\therefore \qquad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 58{\cdot}542$$

The number of degrees of freedom = 10 − 1 = 9, (since we are given 10 frequencies subjected to only one linear constraint $\sum O = \sum E = 10{,}000$).

The tabulated $\chi^2_{0\cdot05}$ for 9 d.f. = 16·919

Since the calculated χ^2 is much greater than the tabulated value, it is highly significant and we reject the null hypothesis. Thus we conclude that the digits are not uniformly distributed in the directory.

Example 13·12. *The following table gives the number of aircraft accidents that occurs during the various days of the week. Find whether the accidents are uniformly distributed over the week.*

Days ...	*Sun.*	*Mon.*	*Tues.*	*Wed.*	*Thus.*	*Fri.*	*Sat.*
No. of accidents ...	*14*	*16*	*8*	*12*	*11*	*9*	*14*

(Given : the values of chi-square significant at 5, 6, 7, d.f. are respectively 11·07, 12·59, 14·07 at the 5% level of significance.

Solution. Here we set up the *null hypothesis that the accidents are uniformly distributed over the week.*

Under the null hypothesis, the expected frequencies of the accidents on each of the days would be :

Days ...	**Sun.**	**Mon.**	**Tues.**	**Wed.**	**Thus.**	**Fri.**	**Sat.**	**Total**
No. of accidents ...	12	12	12	12	12	12	12	84

$$\chi^2 = \frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12}$$

$$= \frac{1}{12}(4 + 16 + 16 + 0 + 1 + 9 + 4) = \frac{50}{12}$$

$$= 4{\cdot}17$$

The number of degrees of freedom

= Number of observations – Number of independent constraints.

$= 7 - 1 = 6$

The tabulated $\chi^2_{0{\cdot}05}$ for 6 d.f. = 12·59

Since the calculated χ^2 is much less than the tabulated value, it is highly insignificant and we accept the null hypothesis. Hence we conclude that the accidents are uniformly distributed over the week.

Example 13·13. *The theory predicts the proportion of beans in the four groups A, B, C and D should be 9 : 3 : 3 : 1. In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory ?* **[Agra Univ. B.Sc., 1991]**

Solution. *Null Hypothesis* : We set up the *null hypothesis that the theory fits well into the experiment, i.e., the experimental results support the theory.*

Under the null hypothesis, the expected (theoretical) frequencies can be computed as follows :

Total number of beans = 882 + 313 + 287 + 118 = 1600

These are to be divided in the ratio 9 : 3 : 3 : 1

$$\therefore \quad E(882) = \frac{9}{16} \times 1600 = 900, \quad E(313) = \frac{3}{16} \times 1600 = 300$$

$$E(287) = \frac{3}{16} \times 1600 = 300, \quad E(118) = \frac{1}{16} \times 1600 = 100$$

$$\therefore \quad \chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$$

$$= \frac{(882-900)^2}{900} + \frac{(313-300)^2}{300} + \frac{(287-300)^2}{300} + \frac{(118-100)^2}{100}$$

$$= 0{\cdot}3600 + 0{\cdot}5633 + 0{\cdot}5633 + 3{\cdot}2400 = 4{\cdot}7266$$

$d.f. = 4 - 1 = 3$, and tabulated $\chi^2_{0{\cdot}05}$ for 3 d.f. = 7·815

Since the calculated value of χ^2 is less than the tabulated value, it is not significant. Hence the null hypothesis may be accepted at 5% level of significance and we may conclude that there is good correspondence between theory and experiment.

Example 13·14. *A survey of 320 families with 5 children each revealed the following distribution :*

No. of boys :	*5*	*4*	*3*	*2*	*1*	*0*
No. of girls :	*0*	*1*	*2*	*3*	*4*	*5*
No. of families :	*14*	*56*	*110*	*88*	*40*	*12*

Is this result consistent with the hypothesis that male and female births are equally probable ?

Solution. Let us set up the *null hypothesis* that *the data are consistent with the hypothesis of equal probability for male and female births.* Then under the null hypothesis :

$$p = \text{Probability of male birth} = \frac{1}{2} = q$$

$$p(r) = \text{Probability of 'r' male births in a family of 5}$$

$$= \binom{5}{r} p^r q^{5-r} = \binom{5}{r} \left(\frac{1}{2}\right)^5$$

The frequency of r male births is given by :

$$f(r) = N.\ p(r) = 320 \times \binom{5}{r} \times \left(\frac{1}{2}\right)^5$$

$$= 10 \times \binom{5}{r} \qquad \ldots(*)$$

Substituting $r = 0, 1, 2, 3, 4$ successively in (*), we get the expected frequencies as follows :

$$f(0) = 10 \times 1 = 10, \qquad f(1) = 10 \times {}^5C_1 = 50$$

$$f(2) = 10 \times {}^5C_2 = 100, \qquad f(3) = 10 \times {}^5C_3 = 100$$

$$f(4) = 10 \times {}^5C_4 = 50, \qquad f(5) = 10 \times {}^5C_5 = 10$$

CALCULATIONS FOR χ^2

Observed Frequencies (O)	*Expected Frequencies (E)*	$(O - E)^2$	$(O - E)^2/E$
14	10	16	1·6000
56	50	36	0·7200
110	100	100	1·0000
88	100	144	1·4400
40	50	100	2·0000
12	10	4	0·4000
Total 320	320		7·1600

$$\therefore \qquad \chi^2 = \sum \left[\frac{(O - E)^2}{E}\right] = 7{\cdot}16$$

Tabulated $\chi^2_{0.05}$ for $6 - 1 = 5$ d.f. is 11·07.

Calculated value of χ^2 is less than the tabulated value, it is not significant at 5% level of significance and hence the null hypothesis of equal probability for male and female births may be accepted.

Example 13·15. *Fit a Poisson distribution to the following data and test the goodness of fit.*

X :	*0*	*1*	*2*	*3*	*4*	*5*	*6*
f :	*275*	*72*	*30*	*7*	*5*	*2*	*1*

Solution. Mean of the given distribution is :

$$\bar{X} = \frac{\sum_i f_i x_i}{N} = \frac{189}{392} = 0{\cdot}482$$

In order to fit a Poisson distribution to the given data, we take the mean (parameter) m of the Poisson distribution equal to the mean of the given distribution, *i.e.,* we take

$$m = \bar{X} = 0{\cdot}482$$

The frequency of r successes is given by the Poisson law as :

$$f(r) = Np(r) = 392 \times \frac{e^{-0{\cdot}482}(0{\cdot}482)^r}{r\,!} \; ; \; r = 0, 1, 2, \ldots, 6$$

Now
$$\begin{aligned} f(0) &= 392 \times e^{-0{\cdot}482} = 392 \times \text{Antilog}\,[-0{\cdot}482 \log e] \\ &= 392 \times \text{Antilog}\,[-0{\cdot}482 \times \log 2{\cdot}7183] \qquad (\because e = 2{\cdot}7183) \\ &= 392 \times \text{Antilog}\,[-0{\cdot}482 \times 0{\cdot}4343] \\ &= 392 \times \text{Antilog}\,[-0{\cdot}2093] \\ &= 392 \times \text{Antilog}\,[\bar{1}{\cdot}7907] = 392 \times 0{\cdot}6176 \\ &= 242{\cdot}1 \end{aligned}$$

$$\begin{aligned} f(1) &= m \times f(0) = 0{\cdot}482 \times 242{\cdot}1 = 116{\cdot}69 \\ f(2) &= \frac{m}{2} \times f(1) = 0{\cdot}241 \times 116{\cdot}69 = 28{\cdot}12 \\ f(3) &= \frac{m}{3} \times f(2) = \frac{0{\cdot}482}{3} \times 28{\cdot}12 = 4{\cdot}518 \\ f(4) &= \frac{m}{4} \times f(3) = \frac{0{\cdot}482}{4} \times 4{\cdot}518 = 0{\cdot}544 \\ (f(5) &= \frac{m}{5} \times f(4) = \frac{0{\cdot}482}{5} \times 0{\cdot}544 = 0{\cdot}052 \\ f(6) &= \frac{m}{6} \times f(5) = \frac{0{\cdot}482}{6} \times 0{\cdot}052 = 0{\cdot}004 \end{aligned}$$

Hence the theoretical Poisson frequencies correct to one decimal place are as given below :

X	0	1	2	3	4	5	6	Total
Expected Frequency	242·1	116·7	28·1	4·5	0·5	0·1	0	392

CALCULATIONS FOR CHI-SQUARE

Observed frequency (O)	*Expected frequency (E)*	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
275	242·1	32·9	1082·41	4·471
72	116·7	44·7	1998·09	17·121
30	28·1	1·9	3·61	0·128

7, 5, 2, 1 } 15	4·5, 0·5, 0·1, 0 } 5·1	9·9	98·01	19·217
392	392·0			40·937

$$\therefore \qquad \chi^2 = \sum \frac{(O-E)^2}{E} = 40{\cdot}937$$

Degrees of freedom = 7 − 1 − 1 − 3 = 2

(One d.f. being lost because of the linear constraint $\sum O = \sum E$; 1 d.f. is lost because the parameter m has been estimated from the given data and is then used for computing the expected frequencies; 3 d.f. are lost because of pooling the last four expected cell frequencies which are less than five.)

Tabulated value of χ^2 for 2 d.f. at 5% level of significance is 5·99.

Conclusion. Since calculated value of χ^2 (40·937) is much greater than 5·99, it is highly significant. Hence we conclude that Poisson distribution is not a good fit to the given data.

EXERCISE 13(b)

1. (*a*) Define Chi-square and obtain its sampling distribution. Mention some prominent features of its frequency curve. Obtain the mean and the variance of the chi-square distribution.

(*b*) Show that the sum of two independent variates having chi-square distributions, has a chi-square distribution.

2. (*a*) Write a short note on the Chi-square test of goodness of fit of a random sample to a hypothetical distribution

(*b*) Describe the Chi-square test of significance and state the various uses to which it can be put.

(*c*) Discuss the χ^2-test of goodness of fit of a theoretical distribution to an observed frequency distribution. How are the degrees of freedom ascertained when some parameters of the theoretical distribution have to be estimated from the data ?

3. (*a*) The following table gives the number of aircraft accidents that occurred during the seven days of the week. Find whether the accidents are uniformly distributed over the week.

Days	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Total
No. of accidents	14	18	12	11	15	14	84

Ans. H_0 : Accidents are uniformly distributed over the week. $\chi^2 = 2{\cdot}143$; Not significant. H_0 may be accepted.

(*b*) A die is thrown 60 times with the following results.

Face	1	2	3	4	5	6
Frequency	8	7	12	8	14	11

Test at 5% level of significance if the die is honest, assuming that $P(\chi^2 > 11{\cdot}1) = 0{\cdot}05$ with 5 d.f. [*Burdwan Univ. B.Sc. (Hons.), 1991*]

4. (*a*) In 250 digits from the lottery numbers, the frequencies of the digits 0, 1, 2, ..., 9 were 23, 25, 20, 23, 23, 22, 29, 25, 33 and 27. Test the hypothesis that they were randomly drawn.

(*b*) 200 digits we chosen at random from a set of tables. The frequencies of the digits were :

Digits	:	0	1	2	3	4	5	6	7	8	9
Frequency	:	18	19	23	21	16	25	22	20	21	15

Use χ^2 test to assess the correctness of hypothesis that the digits were distributed in equal numbers in the table, given that the values of χ^2 are respectively 16·9, 18·3 and 19·7 for 9, 10 and 11 degrees of freedom at 5% level of significance.

[*Delhi Univ. B.Sc., 1992*]

Ans. $\chi^2 = 4{\cdot}3$. Hypothesis seems to be correct.

5. Among 64 offsprings of a certain cross between guinea pigs, 34 were red, 10 were black and 20 were white. According to the genetic model these numbers should be in the ratio 9 : 3 : 4. Are the data consistent with the model at 5 per cent level ?

[You are given that the value of χ^2 with the probability 0·05 being exceeded is 5·99 for 2 d.f. and 3·84 for 1 d.f.]

6. In an experiment on pea-breeding, Mendal obtained the following frequencies of seeds : 315 round and yellow, 101 wrinkled and yellow; 108 round and green, 32 wrinkled and green. Total 556. Theory predicts that the frequencies should be in the proportion 9 : 3 : 3 : 1 respectively. Set up proper hypothesis and test it at 10% level of significance.

Ans. $\chi^2 = 0{\cdot}51$. There seems to be good correspondence between theory and experiment.

7. (*a*) Selfed progenies of a cross between pure strains of plant segregated as follows :

		Early flowering	*Late flowering*
Tall	:	120	48
Short	:	36	13

Do the results agree with the theoretical frequencies which specify a 9 : 3 : 3 : 1 ratio ?

(*b*) Children having one parent of blood-type M and the other type N will always be one of the three types M, MN, N and average proportions of these will be 1 : 2 : 1.

Out of 300 children having one M parent and one N parent, 30% were found to be of type M, 45% of type MN and the remaining of type N. Use χ^2 to test the hypothesis. [*Patna Univ. B.Sc., 1991*]

(*c*) A genetical law says that children having one parent of blood group M and the other parent of blood group N will always be one of the three blood groups M, MN, N ; and that the average number of children in these groups will be in the ratio 1 : 2 : 1. The report on an experiment states as follows : "Of 162 children having one M parent, and one N parent, 28·4% were found to be of group M, 42% of group MN and the rest of the group N". Do the data in the report conform to the expected genetic ratio 1 : 2 : 1 ?

(*d*) A bird watcher sitting in a park has spotted a number of birds belonging to 6 categories. The exact classification is given below:

Category	:	1	2	3	4	5	6
Frequency	:	6	7	13	17	6	5

Test at 5% level of significance whether or not the data is compatible with the assumption that this particular park is visited by birds belonging to these six categories in the proportion 1 : 1 : 2 : 3 : 1 : 1.

[Given $P(\chi^2 \geq 11{\cdot}07) = 0{\cdot}05$ for 5 degrees of freedom]

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1991*]

(*e*) Every clinical thermometer is classified into one of four categories *A*, *B*, *C*, *D* on the basis of inspection and test. From past experience it is known that thermometers produced by a certain manufacturer are distributed among the four categories in the following proportions:

Category	:	*A*	*B*	*C*	*D*
Proportion	:	0·87	0·09	0·03	0·01

A new lot of 1336 thermometers is submitted by the manufacturer for inspection and test and the following distribution into four categories results:

Category	:	*A*	*B*	*C*	*D*
No. of thermometers reported	:	1188	91	47	10

Does this new lot of thermometers differ from the previous experience with regards to proportion of thermometers in each category?

8. (*a*) Five unbiased dice were thrown 96 times and the number of times 4, 5 or 6 was obtained is given below.

No. of dice showing 4, 5 or 6	:	5	4	3	2	1	0
Frequency	:	8	18	35	24	10	1

Fit a suitable distribution and test for the goodness of fit as far as you can proceed without the use of any tables and state how you would proceed further.

[*Gauhati Univ. B.Sc., 1992*]

(*b*) In 120 throws of a single die, the following distribution of faces was obtained:

Faces	:	1	2	3	4	5	6	Total
Frequency	:	30	25	18	10	22	15	120

Compute the statistic you would use to test whether the results constituted refutation of the "equal probability" (null hypothesis). Also state how you would proceed further. [*Nagpur Univ. B.Sc., 1992*]

(*c*) Given below is the number of male and female births in 1,000 families having five children:

Number of male births	*Number of female births*	*Number of families*
0	5	40
1	4	300
2	3	250
3	2	200
4	1	30

Test whether the given data is consistent with the hypothesis that the binomial law holds if the chance of a male birth is equal to that of a female birth.

(*d*)

Five-pig litters Number of males in litter	Number of litters	*Six-pig litters* Number of males in litter	Number of litters
0	2	0	3
1	20	1	16
2	41	2	53
3	35	3	78
4	14	4	53
5	4	5	18
		6	0

Test whether each of the above two samples is a binomial sample (*i*) with $p = 0{\cdot}5$, given *a priori* and (*ii*) with p determined from the data. Test the significance of the difference between the two sample p's.

9. (*a*) The following table gives the count of yeast cells in square of a cyclometer. A square millimeter is divided into 400 equal squares and the number of these squares containing 0, 1, 2, ...cells are recorded—

Number of cells :	0	1	2	3	4	5	6	7	8	9	10
Frequency :	0	20	43	53	86	70	54	37	18	10	5
Number of cells :	11	12	13	14	15	16					
Frequency :	2	2	0	0	0	0					

Fit a Poisson distribution to the data and test the goodness of fit.

(*b*) The following is the distribution of the hourly number of trucks arriving at a company's warehouse.

Trucks arriving per hour :	0	1	2	3	4	5	6	7	8
Frequency :	52	151	130	102	45	12	5	1	2

Find the mean of the distribution and using its mean, (rounded to one decimal) as the parameter λ, fit a Poisson distribution. Test for goodness of fit at the level of significance $\alpha = 0{\cdot}05$

[***Madras Institute of Technology, 1992***]

(*c*) Obtain the equation of the normal curve that may be fitted to the following data :

Class :	60—65	65—70	70—75	75—80	80—85	85—90	90—95	95—100
Frequency :	3	21	150	335	326	135	26	4

Obtain the expected normal frequencies and test the goodness of fit.

10. Aitken gives the following distribution of times shown by two samples of 504 watches each, displayed in watch-maker's windows :

Class interval for time shown	*Frequency of watches from sample I*	*Frequency of watches from sample II*
0—2	75	83
2—4	93	86
4—6	94	94
6—8	76	72
8—10	80	82
10—12	86	87
Total	504	504

Calculate the expected frequencies of watches in the various class intervals under the hypothesis that the times shown are uniformly distributed over the interval (0, 12), separately for the two samples and also for the combined sample of all the 1,008 watches.

Test the goodness of fit for the two samples separately and for the combined sample. Test also the significance of the sum of the values of χ^2 for the two separate samples.

11. The following independent observations were made on the price of grain in 10 consecutive months :

Month :	1	2	3	4	5	6	7	8	9	10
Price (in Rs.) :	115	118	120	140	135	137	139	142	144	150

Test the hypothesis that the expected price in the ith month is Rs. $(100 + 3i)$, $i = 1, 2, \ldots, 10$ with a standard deviation of Rs. 5 under the assumption that the prices are normally distributed.

12. To test a hypothesis H_0, an experiment is performed 3 times. The resulting values of chi-square are 2·37, 1·86 and 3·54, each of which corresponds to one degree of freedom. Show that while H_0 cannot be rejected at 5% level on the basis of any individual experiment, it can be rejected when the three experiments are collectively counted. *[Poona Univ. B.Sc., 1991]*

[**Hint.** Use additive property of chi-square variates.]

13. (*a*) Describe the chi-square test for testing a hypothesis that a normal population has a specified variance σ^2.

(*b*) Give the approximation to the test statistic in (*a*) if n, the sample size, is sufficiently large.

(*c*) A sample of 15 values shows that the s.d. is 6·4. Is this compatible with the hypothesis that the sample is from a normal population with s.d. 5 ?

Ans. $H_0 : \sigma = 5$, $\chi^2 = 24{\cdot}58$; Significant. Population s.d. is not 5.

(*d*) Test the hypothesis that $\sigma = 8$, given that $s = 10$ for a random sample of size 51 from a normal population.

Ans. $Z = \sqrt{2\chi^2} - \sqrt{(2n-1)} = \sqrt{2 \times 79{\cdot}69} - \sqrt{101} = 2{\cdot}57$. Significant at 5% level of significance.

14. (*a*) A manufacturer claims that the life time of a certain brand of batteries produced by his factory has a variance of 5000 (hours)^2. A sample of size 26 has a variance of 7200 (hours)^2. Assuming that it is reasonable to treat these data as a random sample from a normal population, test the manufacturer's claim at the $\alpha = 0{\cdot}02$ level.

Hint. $H_0 : \sigma^2 = 5000$ (hours)^2; $H_1 : \sigma^2 \neq 5000$ (hours)^2 (Two-tailed)

Critical region is : $\chi^2 < \chi^2_{(25)}\,(0{\cdot}99)$ and $\chi^2 > \chi^2_{(25)}\,(0{\cdot}01)$.

(*b*) A manufacturer recorded the cut-off bias (Volt) of a sample of 10 tubes as follows :

12·1, 12·3, 11·8, 12·0, 12·4, 12·0, 12·1, 11·9, 12·2, 12·2

The variability of cut-off bias for tubes of a standard type as measured by the standard deviation is 0·208 volts. Is the variability of the new tube with respect to cut-off bias less than that of the standard type ?

Hint : $H_0 : \sigma^2 = (0.208)^2 \text{ (Volts)}^2 = \sigma^2_0$ (say) ; $H_1 : \sigma^2 < \sigma_0^2$

Critical region is : $\chi^2 < \chi^2_{(n-1)}(1-\alpha) = \chi^2_{(9)}(0.95)$; $\alpha = 0.05$

13·7·3. Independence of Attributes. Let us consider two attributes A and B, A divided into r classes $A_1, A_2, \ldots, A_r$ and B divided into s classes $B_1, B_2, \ldots, B_s$. Such a classification in which attributes are divided into more than two classes is known as *manifold classification*. The various cell frequencies can be expressed in the following table known as $r \times s$ *manifold contingency table* where (A_i) is the number of persons possessing the attribute A_i, $(i = 1, 2, \ldots, r)$, (B_j) is the number of persons possessing the attribute B_j $(j = 1, 2, \ldots, s)$ and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j, $[i = 1, 2, \ldots, r; j = 1, 2, \ldots, s]$. Also

$$\sum_{i=1}^{r} (A_i) = \sum_{j=1}^{s} (B_j) = N, \text{ is the total frequency.}$$

$r \times s$ CONTINGENCY TABLE

A / B	A_1	A_2	......	A_i	......	A_r	*Total*
B_1	(A_1B_1)	(A_2B_1)	......	(A_iB_1)	......	(A_rB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	......	(A_iB_2)	......	(A_rB_2)	(B_2)
⋮ ⋮							
B_j	(A_1B_j)	(A_2B_j)	......	(A_iB_j)	......	(A_rB_j)	(B_j)
⋮ ⋮							
B_s	(A_1B_s)	(A_2B_s)	......	(A_iB_s)	......	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	......	(A_i)	......	(A_r)	N

The problem is to test if two attributes A and B under consideration are independent or not.

Under the *null hypothesis that the attributes are independent*, the theoretical cell frequencies are calculated as follows :

$P[A_i]$ = Probability that a person possesses the attribute A_i

$$= \frac{(A_i)}{N} \; ; \; i = 1, 2, \ldots, r$$

$P[B_j]$ = Probability that a person possesses the attribute B_j

$$= \frac{(B_j)}{N} \; ; \; j = 1, 2, \ldots, s$$

$P[A_iB_j]$ = Probability that a person possesses the attributes A_i and B_j

$$= P(A_i)P(B_j)$$

(By compound probability theorem, since the attributes A_i and B_j are independent, under the null hypothesis).

$$\therefore \quad P[A_iB_j] = \frac{(A_i)}{N} \cdot \frac{(B_j)}{N} \; ; \; i = 1, 2, \ldots, r \; ; \; j = 1, 2, \ldots, s$$

$\therefore$ $(A_iB_j)_0$ = Expected number of persons possessing both the attributes A_i and B_j

$$= N.P[A_iB_j] = \frac{(A_i)(B_j)}{N}$$

$$\Rightarrow \quad (A_iB_j)_0 = \frac{(A_i)(B_j)}{N}, \; (i = 1, 2, \ldots, r \; ; \; j = 1, 2, \ldots, s) \qquad \ldots(13 \cdot 16)$$

By using this formula we can find out expected frequencies for each of the cell-frequencies (A_iB_j), $(i = 1, 2, \ldots, r \; ; \; j = 1, 2, \ldots, s)$, under the null hypothesis of independence of attributes.

The exact test for the independence of attributes is very complicated but a fair degree of approximation is given, for large samples, (large N), by the χ^2-test of goodness of fit, *viz.*,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \left[\frac{\{(A_iB_j) - (A_iB_j)_0\}^2}{(A_iB_j)_0} \right], \qquad \ldots(13 \cdot 16a)$$

which is distributed as a χ^2-variate with $(r - 1)(s - 1)$ d.f. [*c.f.* Note below on degrees of freedom].

Remark. $\phi^2 = \chi^2/N$ is known as *mean-square contingency.*

Since the limits for χ^2 and ϕ^2 vary in different cases, they cannot be used for establishing the closeness of the relationship between qualitative characters under study. Prof. Karl Pearson suggested another measure, known as "*coefficient of mean square contingency*" which is denoted by C and is given by

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{\phi^2}{1 + \phi^2}} \qquad \ldots(13 \cdot 17)$$

Obviously C is always less than unity. The maximum value of C depends on r and s, the number of classes into which A and B are divided. In a $r \times r$ contingency table, the maximum value of $C = \sqrt{(r-1)/r}$. Since the maximum value of C differs for different classification, *viz.*, $r \times r$ $(r = 2, 3, 4, \ldots)$, strictly speaking, the values of C obtained from different types of classifications are not comparable.

Note on Degrees of Freedom (d.f.). The number of independent variates which make up the statistic (*e.g.*, χ^2) is known as the *degrees of freedom* (*d.f.*) and is usually denoted by ν (the letter 'Nu' of the Greek alphabet).

The number of degrees of freedom, in general, is the total number of observations less the number of independent constraints imposed on the observations. For example, if k is the number of independent constraints in a set of data of n observations then $\nu = (n - k)$.

Thus in a set of n observations usually, the degrees of freedom for χ^2 are $(n - 1)$, one d.f. being lost because of the linear constraint $\sum_i O_i = \sum_i E_i = N$, on the frequencies (*c.f.* Theorem 13·3, page 13·12.)

If 'r' independent linear constraints are imposed on the cell frequencies, then the d.f. are reduced by 'r'.

In addition, if any of the population parameter(s) is (are) calculated from the given data and used for computing the expected frequencies then in applying χ^2-test of goodness of fit, we have to subtract one d.f. for each parameter calculated. Thus if 's' is the number of population parameters estimated from the sample observations (n in number), then the required number of degrees of freedom for χ^2-test is $(n - s - 1)$.

If any one or more of the theoretical frequencies is less than 5 then in applying χ^2-test we have also to subtract the degrees of freedom lost in pooling these frequencies with the preceding or succeeding frequency (or frequencies).

In a $r \times s$ contingency table, in calculating the expected frequencies, the row totals, the column totals and the grand totals remain fixed. The fixation of 'r' column totals and 's' row totals imposes $(r + s)$ constraints on the cell frequencies. But since

$$\sum_{i=1}^{r} (A_i) = \sum_{j=1}^{s} (B_j) = N,$$

the total number of independent constraints is only $(r + s - 1)$. Further, since the total number of the cell-frequencies is $r \times s$, the required number of degrees of freedom is :

$$\nu = rs - (r + s - 1) = (r - 1)(s - 1)$$

Example 13·6. *Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the table. Examine whether the nature of the area is related to voting preference in this election.*

Votes for / *Area*	*A*	*B*	*Total*
Rural	*620*	*380*	*1000*
Urban	*550*	*450*	*1000*
Total	*1170*	*830*	*2000*

[*Gujarat Univ. B.Sc., 1990*]

Solution. Under the *null hypothesis* that the nature of the area is independent of the voting preference in the election, we get the observed frequencies as follows :

$$E(620) = \frac{1170 \times 1000}{2000} = 585, \qquad E(380) = \frac{830 \times 1000}{2000} = 415,$$

$$E(550) = \frac{1170 \times 1000}{2000} = 585, \quad \text{and} \quad E(450) = \frac{830 \times 1000}{2000} = 415$$

Aliter. In a 2×2 contingency table, since

$$\text{d.f.} = (2-1)(2-1) = 1,$$

only one of the cell frequencies can be filled up independently and the remaining will follow immediately, since the observed and theoretical marginal totals are fixed. Thus having obtained any one of the theoretical frequencies, (say), $E(620) = 585$, the remaining theoretical frequencies can be easily obtained as follows :

$$E(380) = 1000 - 585 = 415, \; E(550) = 1170 - 585 = 585,$$

and
$$E(450) = 1000 - 585 = 415$$

$$\therefore \qquad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = \frac{(620-585)^2}{585} + \frac{(380-415)^2}{415}$$

$$+ \frac{(550-585)^2}{585} + \frac{(450-415)^2}{415}$$

$$= (35)^2\left[\frac{1}{585} + \frac{1}{415} + \frac{1}{585} + \frac{1}{415}\right]$$

$$= (1225)[2 \times 0{\cdot}002409 + 2 \times 0{\cdot}001709] \doteq 10{\cdot}0891$$

Tabulated $\chi^2_{0\cdot05}$ for $(2-1)(2-1) = 1$ d.f. is 3·841. Since calculated χ^2 is much greater than the tabulated value, it is highly significant and null hypothesis is rejected at 5% level of significance. Thus we conclude that nature of area is related to voting preference in the election.

Example 13·17. (2 × 2 contingency table). *For the* 2×2 *table,*

a	b
c	d

prove that chi-square test of independence gives

$$\chi^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}, \; N = a + b + c + d \qquad \ldots(13{\cdot}18)$$

[*Gauhati Univ. B.Sc., 1992*]

Solution. Under the hypothesis of independence of attributes,

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

$$E(a) = \frac{(a+b)(a+c)}{N}$$

$$E(b) = \frac{(a+b)(b+d)}{N}$$

$$E(c) = \frac{(a+c)(c+d)}{N}$$

and $$E(d) = \frac{(b+d)(c+d)}{N}$$

$$\therefore \quad \chi^2 = \frac{[a-E(a)]^2}{E(a)} + \frac{[b-E(b)]^2}{E(b)} + \frac{[c-E(c)]^2}{E(c)} + \frac{[d-E(d)]^2}{E(d)} \qquad \ldots(*)$$

$$a - E(a) = a - \frac{(a+b)(a+c)}{N}$$

$$= \frac{a(a+b+c+d) - (a^2+ac+ab+bc)}{N} = \frac{ad-bc}{N}$$

Similarly, we will get

$$b - E(b) = -\frac{ad-bc}{N} = c - E(c)\,;\; d - E(d) = \frac{ad-bc}{N}$$

Substituting in (*), we get

$$\chi^2 = \frac{(ad-bc)^2}{N^2}\left[\frac{1}{E(a)} + \frac{1}{E(b)} + \frac{1}{E(c)} + \frac{1}{E(d)}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\left\{\frac{1}{(a+b)(a+c)} + \frac{1}{(a+b)(b+d)}\right\} + \left\{\frac{1}{(a+c)(c+d)} + \frac{1}{(b+d)(c+d)}\right\}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\frac{b+d+a+c}{(a+b)(a+c)(b+d)} + \frac{b+d+a+c}{(a+c)(c+d)(b+d)}\right]$$

$$= (ad-bc)^2\left[\frac{c+d+a+b}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$= \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Example 13·18. *A random sample of students of Bombay University was selected and asked their opinions about 'autonomous colleges'. The results are given below. The same number of each sex was included within each class-group. Test the hypothesis at 5% level that opinions are independent of the class groupings :—*

Class	*Numbers*		*Total*
	Favouring 'autonomous colleges'	*Opposed to 'autonomous colleges'*	
F.Y. B.A./B.Sc./B.Com.	*120*	*80*	*200*
S.Y. B.A./B.Sc./B.Com.	*130*	*70*	*200*
T.Y. B.A./B.Sc./B.Com.	*70*	*30*	*100*
M.A./M.Sc./M.Com.	*80*	*20*	*100*
Total	*400*	*200*	*600*

[***Bombay Univ. B.Sc., April 1989***]

Solution. We set up the null hypothesis that the opinions about autonomous colleges are independent of the class-groupings.

Here the frequencies are arranged in the form of a 4×2 contingency table. Hence the d.f. are $(4-1) \times (2-1) = 3 \times 1 = 3$. Hence we need to compute independently only three expected frequencies and the remaining expected frequencies can be obtained by subtraction from the row and column totals.

Under the null hypothesis of independence :

$$E(120) = \frac{400 \times 200}{600} = 133 \cdot 33 \ ; \ E(130) = \frac{400 \times 200}{600} = 133 \cdot 33$$

$$E(70) = \frac{400 \times 100}{600} = 66 \cdot 67$$

Now the table of expected frequencies can be completed as shown below :

Class	*Number*		*Total*
	Favouring 'autonomous colleges'	*Opposed to 'autonomous colleges'*	
F.Y.B.A./B.Sc./B.Com.	133·33	200 − 133·33 = 66·67	200
S.Y.B.A./B.Sc./B.Com.	133·33	200 − 133·33 = 66·67	200
T.Y.B.A./B.Sc./B.Com.	66·67	100 − 66·67 = 33·33	100
M.A./M.Sc./M.Com.	66·67	100 − 66·67 = 33·33	100
Total	400	200	600

CALCULATIONS FOR CHI-SQUARE

O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
120	133·33	−13·33	177·6889	1·3327
130	133·33	−3·33	11·0889	0·0832
70	66·67	3·33	11·0889	0·1663
80	66·67	13·33	177·6889	2·6652
80	66·67	13·33	177·6889	2·6652
70	66·67	3·33	11·0889	0·1663
30	33·33	−3·33	11·0889	0·3327
20	33·33	−13·33	177·6889	5·3312
Total 400	400			12·7428

$$\therefore \qquad \chi^2 = \sum \frac{(O-E)^2}{E} = 12{\cdot}7428$$

Tabulated (critical) value of χ^2 for $(4-1)\times(2-1) = 3$ d.f. at 5% level of significance is 7·815.

Conclusion. Since calculated value of χ^2 is greater than the tabulated value, it is significant at 5% level of significance and we reject the null hypothesis. Hence, we conclude that the opinions about autonomous colleges' are dependent on the class-groupings.

Example 13·19. *Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows :*

Researcher	*No. of students in each level* Below Average	*Average*	*Above Average*	*Genius*	*Total*
X	*86*	*60*	*44*	*10*	*200*
Y	*40*	*33*	*25*	*2*	*100*
Total	*126*	*93*	*69*	*12*	*300*

Would you say that the sampling techniques adopted by the two researchers are significantly different ? (Given 5% value of χ^2 for 2 d.f. and 3 d.f. are 5·991 and 7·82 respectively.)

Solution. We set up the null hypothesis that the data obtained are independent of the sampling techniques adopted by the two researchers. In other words, the null hypothesis is that their is no significant difference between the sampling techniques used by the two researchers for collecting the required data.

Here we have a 4 × 2 contingency table and $d.f. = (4-1) \times (2-1) = 3 \times 1 = 3$. Hence we need to compute only 3 independent expected frequencies and the remaining expected frequencies can be obtained by subtraction from the marginal row and column totals.

Under the null hypothesis of independence, we have

$$E(86) = \frac{126 \times 200}{300} = 84; \; E(60) = \frac{93 \times 200}{300} = 62;$$

$$E(44) = \frac{69 \times 200}{300} = 46$$

The table of expected frequencies can now be completed as shown below :

Researcher	No. of students in each level				Total
	Below Average	Average	Above average	Genius	
X	84	62	46	200 − 192 = 8	200
Y	126 − 84 = 42	93 − 62 = 31	69 − 46 = 23	12 − 8 = 4	100
Total	126	93	69	12	300

Since we cannot apply the χ^2-test straightway here as the last frequency is less than 5, we should use the technique of pooling in this case as given below :

CALCULATIONS FOR CHI-SQUARE

O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
86	84	2	4	0·048
60	62	−2	4	0·064
44	46	−2	4	0·087
10	8	2	4	0·500
40	42	−2	4	0·095
33	31	2	4	0·129
$\left.\begin{matrix}25\\2\end{matrix}\right\}27$	$\left.\begin{matrix}23\\4\end{matrix}\right\}27$	0	0	0
Total 300	300	0	0	0·923

After pooling, $\chi^2 = \sum \left[\frac{(O-E)^2}{E}\right] = 0{\cdot}923$

and the $d.f. = (4-1) \times (2-1) - 1 = 3 - 1 = 2$, since 1 $d.f.$ is lost in the method of pooling.

Tabulated value of χ^2 for 2$d.f.$ at 5% level of significance is 5·991.

Conclusion. Since calculated value is less that the tabulated value, null hypothesis may be accepted at 5% level of significance and we may conclude that there is no significant difference in the sampling techniques used by the two researchers.

13·8. Yates' Correction. In a 2 × 2 contingency table, the number of *d.f.* is (2 – 1) (2 – 1) = 1. If any one of the theoretical cell frequencies is less than 5, then the use of pooling method for χ^2-test results in χ^2 with 0 *d.f.* (since 1 *d.f.* is lost in pooling) which is meaningless. In this case we apply a correction due to F. Yates (1934), which is usually known as *"Yates' Correction for Continuity"*. [As already pointed out, χ^2 is a continuous distribution and it fails to maintain its character of continuity if any of the expected frequency is less than 5; hence the name 'Correction for Continuity']. This consists in adding 0·5 to the cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly. The χ^2-test of goodness of fit is then applied without pooling method.

For a 2 × 2 contingency table,

a	b
c	d

, we have

$$\chi^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

According to Yate's correction, as explained above, we subtract (or add) $\frac{1}{2}$ from a and d and add (subtract) $\frac{1}{2}$ to b and c so that the marginal totals are not disturbed at all. Thus, corrected value of χ^2 is given as

$$\chi^2 = \frac{N\left[(a \mp \frac{1}{2})(d \mp \frac{1}{2}) - (b \pm \frac{1}{2})(c \pm \frac{1}{2})\right]^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$\text{Numerator} = N\left[(ad - bc) \mp \tfrac{1}{2}(a+b+c+d)\right]^2$$

$$= N\left[\,|\,ad - bc\,| - \frac{N}{2}\right]^2$$

$$\therefore \quad \chi^2 = \frac{N\,[\,|\,ad - bc\,| - N/2]^2}{(a+c)(b+d)(a+b)(c+d)} \qquad \ldots(13\cdot18a)$$

Remarks 1. If N is large, the use of Yate's correction will make very little difference in the value of χ^2. If, however, N is small, the application of Yates' correction may overstate the probability.

2. It is recommended by many authors and it seems quite logical in the light of the above discussion that Yates' correction be applied to every 2 × 2 table, even if no theoretical cell frequency is less than 5.

13·9. Brandt and Snedecor Formula for 2 × k Contingency Table. Let the observations a_{ij}, $(i = 1, 2; j = 1, 2, \ldots, k)$ be arranged in a $2 \times k$ contingency table as follows :

A / B	A_1	A_2	......	A_i	......	A_k	Total
B_1	a_{11}	a_{12}	......	a_{1i}	......	a_{1k}	m_1
B_2	a_{21}	a_{22}	......	a_{2i}	......	a_{2k}	m_2
Total	n_1	n_2	......	n_i	......	n_k	N

Under the hypothesis of independence of attributes, we have

$$E(a_{1i}) = \frac{n_i \times m_1}{N}\ ;\ E(a_{2i}) = \frac{n_i \times m_2}{N};\quad i = 1, 2, \ldots, k$$

$$\therefore \qquad \chi^2 = \sum_{i=1}^{k}\left[\frac{\{a_{1i} - E(a_{1i})\}^2}{E(a_{1i})} + \frac{\{a_{2i} - E(a_{2i})\}^2}{E(a_{2i})}\right]$$

$$= \sum_{i=1}^{k}\left\{\frac{\left(a_{1i} - \frac{m_1 n_i}{N}\right)^2}{\left(\frac{m_1 n_i}{N}\right)} + \frac{\left(a_{2i} - \frac{m_2 n_i}{N}\right)^2}{\left(\frac{m_2 n_i}{N}\right)}\right\}$$

$$= \sum_{i=1}^{k}\left[\frac{N n_i}{m_1}\left(\frac{a_{1i}}{n_i} - \frac{m_1}{N}\right)^2 + \frac{N n_i}{m_2}\left(\frac{a_{2i}}{n_i} - \frac{m_2}{N}\right)^2\right]$$

$$= \sum_{i=1}^{k}\left[\frac{n_i}{p}\left(p_i - p\right)^2 + \frac{n_i}{q}\left(q_i - q\right)^2\right]$$

where $\qquad p_i = \frac{a_{1i}}{n_i},\ q_i = 1 - p_i = \frac{a_{2i}}{n_i}$

and $\qquad p = \frac{m_1}{N},\ q = 1 - p = \frac{m_2}{N}$

$$\therefore \qquad \chi^2 = \sum_{i=1}^{k}\left[\frac{n_i}{p}\left(p_i - p\right)^2 + \frac{n_i}{q}\left\{\left(1 - p_i\right) - \left(1 - p\right)\right\}^2\right]$$

$$= \sum_{i=1}^{k} n_i \left(p_i - p\right)^2 \left\{\frac{1}{p} + \frac{1}{q}\right\}$$

$$= \sum_{i=1}^{k} n_i \left(p_i - p\right)^2 \frac{1}{pq} \qquad [\because p + q = 1]$$

$$= \frac{1}{pq}\sum_{i=1}^{k} n_i \left(p_i^2 + p^2 - 2p_i\, p\right)$$

$$= \frac{1}{pq} \left[\sum_{i=1}^{k} n_i p_i^2 + p^2 \sum_{i=1}^{k} n_i - 2p \sum_{i=1}^{k} p_i n_i \right]$$

$$= \frac{1}{pq} \left[\sum_{i=1}^{k} n_i p_i^2 + p^2 N - 2p . Np \right]$$

$$= \frac{1}{pq} \left[\sum_{i=1}^{k} n_i p_i^2 - Np^2 \right]$$

But $$\sum_{i=1}^{k} n_i p_i^2 = \sum_{i=1}^{k} n_i p_i . p_i = \sum_{i=1}^{k} a_{1i} p_i$$

$$\therefore \quad \chi^2 = \frac{1}{pq} \left[\sum_{i=1}^{k} a_{1i} p_i - Np^2 \right] = \frac{1}{pq} \left[\sum_{i=1}^{k} \frac{a_{1i}^2}{n_i} - Np^2 \right] \quad \ldots(13·19)$$

$$= \frac{1}{pq} \left[\sum_{i=1}^{k} \frac{a_{1i}^2}{n_i} - m_1 p \right] = \frac{1}{pq} \left[\sum_{i=1}^{k} \frac{a_{1i}^2}{n_i} - \frac{m_1^2}{N} \right] \quad \ldots[13·19(a)]$$

Example 13·20. *The following table shows three age groups of boys and girls, (a) the number of children affected by a non-infectious disease and (b) the total number of children exposed to risk.*

	Boys			*Girls*		
	I	*II*	*III*	*I*	*II*	*III*
(a)	*60*	*25*	*48*	*96*	*18*	*42*
(b)	*240*	*470*	*350*	*530*	*200*	*210*

(i) Test whether there are differences between the incidence rates in the three age groups of boys.

(ii) Test whether the boys and girls are equally susceptible or not.

Solution. (*i*) We set up the *null hypothesis* (H_0) *that there is no significant difference between the incidence rates in the three age-groups of boys.* In the notations of § 13·9, we have

$$\left.\begin{array}{l} a_{11} = 60,\ a_{12} = 25,\ a_{13} = 48,\ m_1 = 133 \\ n_1 = 240,\ n_2 = 470,\ n_3 = 350,\ N = 1060 \end{array}\right\}$$

$$p = \frac{m_1}{N} = \frac{133}{1060} = 0·1255,\ q = 1 - p = 0·8745$$

Substituting these values in (13·19*a*), we get

$$\chi^2 = \frac{1}{(0·1255)(0·8745)} \left[15·00 + 1·33 + 6·58 - 133 \times 0·1255 \right]$$

$$= \frac{6·2187}{0·1097} = 56·688$$

Here $\nu = (3-1)(2-1) = 2$.

The tabulated value of χ^2 for 2 degrees of freedom at 5% level of significance is 5·991. Since calculated value of χ^2 is much greater than tabulated value, we reject the null hypothesis and conclude that the incidence rates in the three age-groups of boys differ significantly.

(*ii*) Here we set up the *null hypothesis that the boys and girls are equally susceptible to the disease.* In the usual notations, we have

$$a_{11} = 60 + 25 + 48 = 133 \text{ and } a_{12} = 96 + 18 + 42 = 156, m_1 = 289$$

$$n_1 = 240 + 470 + 350 = 1060 \text{ and } n_2 = 530 + 200 + 210 = 940, N = 2000$$

$$\therefore \quad p = \frac{289}{2000} = 0{\cdot}1445, \quad q = 1 - p = 0{\cdot}8555$$

$$Np^2 = 289 \times 0{\cdot}1445 = 41{\cdot}76$$

$$\therefore \quad \chi^2 = \frac{1}{(0{\cdot}1445)\,(0{\cdot}8555)}\left[16{\cdot}69 + 25{\cdot}89 - 41{\cdot}76\right] = 6{\cdot}605$$

Here $\quad v = (2-1)\,(2-1) = 1$

From the tables, the value of χ^2 for 1 degree of freedom at 5% level of significance is 3·841 which is much less than the calculated value. We, therefore, reject the null hypothesis and conclude that boys and girls are not equally susceptible to the disease.

Example 13·21. *Two samples of sizes N_1, N_2 have respectively frequencies $f_1, f_2, \ldots, f_n$ and $f_1', f_2', \ldots, f_n'$ under the same headings. Show that χ^2 for such a distribution is equal to*

$$\sum_{r=1}^{n} N_1 N_2 \left[\frac{\left(\dfrac{f_r}{N_1} - \dfrac{f_r'}{N_2}\right)^2}{f_r + f_r'}\right]$$

[*Allahabad Univ. B.Sc., 1992*]

Solution. The $2 \times n$ contingency table for which χ^2 is to be calculated is given below :

B \ A	A_1	A_2	...	A_r	...	A_n	Total
B_1	f_1	f_2	...	f_r	...	f_n	N_1
B_2	f_1'	f_2'	...	f_r'	...	f_n'	N_2

Under the hypothesis of independence of attributes, we have

$$E(f_r) = \frac{N_1(f_r + f_r')}{N_1 + N_2}, \quad E(f_r') = \frac{N_2(f_r + f_r')}{N_1 + N_2}$$

$$\chi^2 = \sum_{r=1}^{n}\left[\frac{\{f_r - E(f_r)\}^2}{E(f_r)} + \frac{\{f_r' - E(f_r')\}^2}{E(f_r')}\right]$$

$$= \sum_{r=1}^{n} \left[\frac{(N_1+N_2)}{N_1(f_r+f_r')} \cdot \left\{ f_r - \frac{N_1(f_r+f_r')}{N_1+N_2} \right\}^2 + \frac{(N_1+N_2)}{N_2(f_r+f_r')} \left\{ f_r' - \frac{N_2(f_r+f_r')}{N_1+N_2} \right\}^2 \right]$$

$$= \sum_{r=1}^{n} \left[\frac{(N_2 f_r - N_1 f_r')^2}{N_1(N_1+N_2)(f_r+f_r')} + \frac{(N_1 f_r' - N_2 f_r)^2}{N_2(N_1+N_2)(f_r+f_r')} \right]$$

$$= \sum_{r=1}^{n} \frac{(N_2 f_r - N_1 f_r')^2}{(N_1+N_2)(f_r+f_r')} \left[\frac{1}{N_1} + \frac{1}{N_2} \right]$$

$$= \sum_{r=1}^{n} \left[\frac{N_1 N_2}{f_r+f_r'} \left(\frac{f_r}{N_1} - \frac{f_r'}{N_2} \right)^2 \right]$$

EXERCISE 13(c)

1. (*a*) What is contingency table ? Describe how the χ^2 distribution may be used to test whether the two criteria of classification in an $m \times n$ contingency table are independent.

(*b*) State the hypothesis you test using the Chi-square statistic in a contingency table.

(*c*) Describe the χ^2 test for independence of attributes, stating clearly the conditions for the validity. Give a rule for calculating the number of degrees of freedom to be assigned to χ^2. Illustrate your answer with an $m \times n$ contingency table explaining the null hypothesis that is being tested.

2. Of 'A' candidates taking a certain paper, 'a' are successful, of 'B' taking another paper, 'b' are successful. Show how the significance of the difference between the ratios a/A, b/B may be tested (*i*) by a χ^2 test on contingency table and (*ii*) by comparing the difference with its standard error assessed by means of a binomial distribution. (You may assume all frequencies are sufficiently large.) Prove algebraically that the value of χ^2 is the square of the ratio of $(a/A - b/B)$ to its standard error.

3. (*a*) Show that for the entries in the following $2 \times r$ contingency table,

	A_1	A_2		A_i		A_r	Total
B_1	a_1	a_2	......	a_i	......	a_r	a
B_2	b_1	b_2	......	b_i	......	b_r	b
Total	n_1	n_2	......	n_i	......	n_r	n

the value of χ^2 is

$$\chi^2 = \sum_{i=1}^{r} \omega_i (p_i - p)^2$$

where $p_i = \frac{a_i}{n_i}, p = \frac{a}{n}, \ \omega_i = \frac{n_i}{pq}$ and $q_i = 1 - p_i, \ q = \frac{b}{n}$

[*Madurai Univ. B.Sc., Oct. 1988*]

(*b*) Given χ^2 contingency table representing two independent samples :

					Total
Sample I	μ_1	μ_2	...	μ_r	m
Sample II	v_1	v_2	...	v_r	n
Total	$\mu_1 + v_1$	$\mu_2 + v_2$	...	$\mu_r + v_r$	$m + n$,

show that

$$\chi^2 = \frac{1}{\omega\,(1-\omega)}\left[\sum_{i=1}^{r} \mu_i \omega_i - m\,\omega\right]$$

where $\omega_i = \dfrac{\mu_i}{\mu_i + v_i}$ and $\omega = \dfrac{m}{m+n}$,

can be used to test whether the samples are drawn from the sample population. Clearly state the underlying assumptions, and give the number of degrees of freedom.

(*c*) In a 2×3 contingency table if $N = x + y + z, N' = x' + y' + z'$ and $N = N'$, show that

$$\chi^2 = \frac{(x-x')^2}{x+x'} + \frac{(y-y')^2}{y+y'} + \frac{(z-z')^2}{z+z'}$$

(*Poona Univ. B.Sc., 1990*)

(*d*) Show that for a 2×2 table, the value of χ^2, after applying Yates' correction for continuity is

$$\frac{N}{D}\left(ad - bc - \frac{N}{2}\right)^2 \quad \text{or} \quad \frac{N}{D}\left(ad - bc + \frac{N}{2}\right)^2$$

according as $ad - bc > 0$ or < 0 respectively, where

$$D = (a+b)\,(a+c)\,(b+d)\,(c+d).$$

(*e*) What is Yates' correction ? Show that for a 2×2 contingency table, the value of χ^2 after applying this correction is :

$$\chi^2 = \frac{N[\,|ad - bc| - N/2]^2}{(a+b)\,(a+c)\,(b+d)\,(c+d)}$$

[*Marathwada Univ. M.Sc., 1991*]

4. Consider the following 2×2 table of observed frequencies based on random samples (with replacement) of sizes $n_{.1}$ and $n_{.2}$ from two populations :

	Population I	*Population II*	*Total*
Class A	n_{11}	n_{12}	$n_{1.}$
Class B	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

(*i*) Define the χ^2-statistic to be used for test of homogeneity of the two populations.

(*ii*) Show that

$$\chi^2 = \frac{n(n_{11}\,n_{22} - n_{12}n_{21})^2}{(n_{1.}\,n_{.1}\,n_{2.}\,n_{.2})}$$

(*iii*) Let

$$u = \frac{n_{11}}{n_{.1}} - \frac{n_{12}}{n_{.2}}$$

Calculate the mean and variance of u and indicate how you may estimate them.

5. (*a*) In an epidemic of certain disease 92 children contracted the disease. Of these, 41 received no treatment and of these 10 showed after-effects. Of the remainder who did receive treatment, 17 showed after-effects. Test the hypothesis that treatment was not effective.

(*b*) Can vaccination be regarded as a preventive measure of small-pox as evidenced by the following data ?

"Of 1482 persons exposed to smallpox in a locality, 368 in all were attacked. Of these 1482 persons, 343 were vaccinated and of these, only 35 were attacked".

6. (*a*) Define χ^2. Cite some statistical problems where you can apply χ^2 for testing statistical hypothesis.

In an experiment on immunization of cattle from tuberculosis the following results were obtained :

	Affected	*Unaffected*
Inoculated	12	28
Not inoculated	13	7

Examine the effect of vaccine in controlling the incidence of the disease.

(*b*) What are contingency tables ? What is tested there ? Explain the test procedure therein.

The following data is collected on two characters :

	Cinegoers	*Non-cinegoers*
Literate	83	57
Illiterate	45	68

Based on this, can you conclude that there is no relation between the habit of cinema going and literacy ?

7. (*a*) To find whether a certain vaccination prevents a certain disease or not, an experiment was conducted and the following figures in various classes were obtained, A showing vaccination and B attacked by the disease.

	A	α	Total
B	69	10	79
β	91	30	121
Total	160	40	200

Using χ^2-test, analyse the results of the experiment for independence between A and B; examine whether Yate's correction modifies the conclusion or not. Test also the significance of the difference between the proportions of persons attacked by the disease among vaccinated and non-vaccinated which are 69/160 and 10/40.

(*b*) A theory in finance known as Random Walk Theory suggests that short term changes in stock prices follow a random pattern. According to this theory, yesterday's price change can tell us virtually nothing of value about to-day's price change. Let us denote the change in price of a stock on day t by ΔP_t and the change on the next day by ΔP_{t+1}. Suppose we observe price changes of 240 stocks that have been randomly selected and obtain the results shown in the table below :

	$\Delta P_t > 0$	$\Delta P_t \le 0$	*Total*
$\Delta P_{t+1} > 0$	47	53	100
$\Delta P_{t+1} \le 0$	63	77	140
Total	110	130	240

Test the hypothesis that the change in stock price on day $(t+1)$ is independent of that on day t.

[Delhi Univ. M.A. (Eco.), 1987]

8. (*a*) Show that the value of χ^2 for 2×2 contingency table

a	b
c	d

is $$\chi^2 = \frac{N(ad - bc)^2}{(a + c)\,(b + d)\,(a + b)\,(c + d)},$$

where $N = a + b + c + d$.

(*b*) Let X and Y denote the number of successes and failures respectively in n independent Bernoulli trials with p as the probability of success in each trial. Show that

$$\frac{(X - np)^2}{np} + \frac{[Y - n(1-p)]^2}{n(1 - p)}$$

can be approximated by a chi-square distribution with one degree of freedom when n is large. *[Delhi Univ. M.A. (Eco.), 1986]*

9. Show that for $r \times s$ contingency table :

(*a*) Number of degrees of freedom is $(r-1) \times (s-1)$

(*b*) $\chi^2 = N(s-1)$ or $\chi^2 = N(r-1)$, whichever is less

(*c*) $E(\chi^2) = N(r-1)(s-1)/(N-1)$

(*d*) $\max(C) = [(s-1)/s]^{1/2}$, $r = s$,

where C is the coefficient of contingency and N is the total frequency.

10. (*a*) 1072 college students were classified according to their intelligence and economic conditions. Test whether there is any association between intelligence and economic conditions.

		Intelligence			
		Excellent	*Good*	*Mediocre*	*Dull*
Economic Conditions	*Good*	48	199	181	82
	Not good	81	185	190	106

(*b*) Below is given the distribution of hair colours for either sex in a university :

Hair colour	(1) Fair	(2) Red	(3) Medium	(4) Dark	(5) Jet black	Total
Boys	592	119	849	504	36	2100
Girls	544	97	677	451	14	1783
Total	1136	216	1526	955	50	3883

Test the homogeneity of hair colour for either sex. If the result is significant at 5 per cent level, explain the reason why it should be so.

11. (*a*) The following data are for a sample of 300 car owners who were classified with respect to age and the number of accidents they had during the past two years. Test whether there is any relationship between these two variables.

		Accidents	
	0	1 or 2	3 or more
Age ≤ 21	8	23	14
Age 22 — 26	21	42	12
Age ≥ 27	71	90	19

(*b*) For the data in the following table, test for independence between a person's ability in Mathematics and interest in Economics.

		Ability in Mathematics		
		Low	*Average*	*High*
Interest in Economics	Low	63	42	15
	Average	58	61	31
	High	14	47	29

State clearly the assumptions underlying your test procedure.

[*Delhi Univ. M.A. (Eco.), 1988*]

12. The following table gives for a sample of married women, the level of education and marriage adjustment score :

		Marriage—adjustment score			
		Very low	*Low*	*High*	*Very high*
Level of Education	College	24	97	62	58
	High school	22	28	30	41
	Middle School	32	10	11	20

Can you conclude from the above, 'the higher the level of education, the greater is the degree of adjustment in marriage' ?

13. (*a*) The table below shows results of a survey in which 250 respondents were categorized according to level of education and attitude towards students' demonstrations at a certain college. Test the hypothesis that the two criteria of classification are independent. Let $\alpha = 0{\cdot}05$.

	Attitude		
Education	*Against*	*Neutral*	*For*
Less than high school	40	25	5
High school	40	20	5
Some college	30	15	30
College graduate	15	15	10

(*b*) Test the hypothesis that there is no difference in the quality of the four kinds of tyres *A*, *B*, *C* and *D* based on the data given below. Use 5% level of significance.

	Tyre Brand			
	A	*B*	*C*	*D*
Failed to last 40,000 kms.	26	23	15	32
Lasted from 40,000 kms. to 60,000 kms.	118	93	116	121
Lasted more than 60,000 kms.	56	84	69	47

[***Bangalore Univ. B.E., 1992***]

(*c*) The results of a survey regarding radio listeners' preference for different types of music are given in the following table, with listeners classified by age group. Is preference of type of music influenced by age ?

Type of music	*Age group*		
preferred	*19—25*	*26—35*	*Above 36*
National music	80	60	9
Foreign music	210	325	44
Indifferent	16	45	132

14. (*a*) If $x_1, x_2, \ldots, x_k$ represent the respective number of successes in k samples each of n trials, by considering a suitable $2 \times k$ contingency table, derive an expression for χ^2 to test the homogeneity of this data.

(*b*) It was decided to check the dental health of children in 8 districts of a town. The condition of the teeth of 36 children from each district was examined and classified as either good or poor. The number of children with teeth in a poor condition from each of the districts was 9, 14, 12, 18, 7, 10, 15, 11. Can it be concluded that the dental health of children does not vary between districts ?

13·9·1. χ^2-test of Homogeneity of Correlation Coefficients. Let $r_1, r_2, \ldots, r_k$ be k estimates of correlation coefficients from independent samples of sizes $n_1, n_2, \ldots, n_k$ respectively.

We want to test the hypothesis that these sample correlation coefficients are the estimates of the same correlation coefficient ρ from a bivariate normal population.

Obtain the values of $z_1, z_2, \ldots, z_k$ from the Table of Fisher' z - transformation or from

$$z_i = \frac{1}{2} \log_e \left(\frac{1 + r_i}{1 - r_i} \right) = \tanh^{-1} r_i \,; \; i = 1, 2, \ldots, k \qquad \ldots(13\cdot20)$$

These z_i's are normally distributed about a common mean

$$\xi = \frac{1}{2} \log_e \left(\frac{1 + \rho}{1 - \rho} \right) \text{ and variance} = \frac{1}{n_i - 3} \qquad \ldots(13\cdot21)$$

The minimum variance estimate z of the common mean ξ of Z's is obtained by weighting the values z_i's inversely with their respectively variances. The estimate of z is, therefore,

$$\bar{z} = \frac{\sum_i z_i (n_i - 3)}{\sum_i (n_i - 3)} \qquad \text{(c.f. § 14·7·2)}$$

so that $(z_i - \bar{z}) \sqrt{n_i - 3}\,;\; i = 1, 2, \ldots, k$ are independent standard normal variates. Hence $\sum\limits_{i=1}^{k} (n_i - 3)(z_i - \bar{z})^2$ is a χ^2-variate with $(k - 1)$ d.f. [By additive property of χ^2-distribution, one d.f. being lost since z has been determined from the data.]

If χ^2 value thus obtained is greater than 5 per cent value of χ^2 for $(k - 1)$ d.f., the hypothesis of homogeneity of correlation coefficients is rejected. If not, the correlation coefficients are supposed to be homogeneous in which case we combine the sample correlation coefficients to find the estimate $\hat{\rho}$ of the population correlation coefficient ρ.

We have $\qquad \bar{z} = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right)$

$$\Rightarrow \qquad (1 + \hat{\rho}) = (1 - \hat{\rho})\, e^{2\bar{z}}$$

$$\Rightarrow \qquad (1 + e^{2\bar{z}})\, \hat{\rho} = e^{2\bar{z}} - 1$$

$$\Rightarrow \qquad \hat{\rho} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1} = \tanh \bar{z} \qquad \ldots(13\cdot22)$$

Remark. For testing the homogeneity of independent estimates of the parent partial correlation coefficient, the above formulae hold with the only difference that for a partial correlation coefficient of order s, n_i will be replaced by $n_i - s$.

Example 13·22. *The correlation coefficient between daily ration of green grass and rate of growing calves on the basis of observations taken on 10, 14, 16, 20, 25 and 28 cows at six farms were found to be 0·318, 0·106, 0·253, 0·340, 0·116 and 0·112. Can these be considered homogeneous ? If so, estimate the common correlation coefficient.*

Solution. H_0 : *The given values of sample correlation coefficients are homogeneous* or the samples are from equally correlated populations.

Using (13·20), we get

$$z_1 = 0{\cdot}3294, \quad z_2 = 0{\cdot}1063, \quad z_3 = 0{\cdot}2586$$
$$z_4 = 0{\cdot}3541, \quad z_5 = 0{\cdot}1165, \quad \text{and} \quad z_6 = 0{\cdot}1125$$

$$\therefore \qquad \bar{z} = \sum_i z_i (n_i - 3) / \sum_i (n_i - 3) = 0{\cdot}1919$$

Now $\qquad \chi^2 = \Sigma (n_i - 3)(z_i - \bar{z})^2 = 0{\cdot}1008$

Tabulated value of χ^2 for $(6 - 1) = 5$, degrees of freedom at 5% level of significance is 11·070.

Since the calculated value is less than the tabulated value, we may accept the null hypothesis that the sample correlation coefficients are homogeneous.

If $\hat{\rho}$ is the pooled estimate of the population correlation coefficient, then using (13·22), we get

$$\hat{\rho} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1} = \frac{1{\cdot}468 - 1}{1{\cdot}468 + 1} = 0{\cdot}1894$$

13·10. Bartlett's Test for Homogeneity of Several Independent Estimates of the Same Population Variance. Let

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2, \quad (i = 1, 2, \ldots, k)$$

be the unbiased estimate of the population variance, obtained from the ith sample X_{ij}, $(j = 1, 2, \ldots, n_i)$ and based on $\nu_i = (n_i - 1)$ degrees of freedom, all the k samples being independent.

Under the null hypothesis that the samples come from the same population with variance σ^2, *i.e.*, the independent estimates S_i^2, $(i = 1, 2, \ldots, k)$ of σ^2 are homogeneous, Bartlett proved that the statistic

$$\chi^2 = \sum_{i=1}^{k} \left(\nu_i \log \frac{S^2}{S_i^2} \right) \Bigg/ \left[1 + \frac{1}{3(k-1)} \left\{ \sum_i \left(\frac{1}{\nu_i} \right) - \frac{1}{\nu} \right\} \right] \quad \ldots(13{\cdot}23)$$

where $\quad S^2 = \dfrac{\Sigma \nu_i S_i^2}{\Sigma \nu_i} = \dfrac{\Sigma \nu_i S_i^2}{\nu}, \quad \sum_{i=1}^{k} \nu_i = \nu \qquad \ldots(*)$

follows chi-square distribution with $(k - 1)$ degrees of freedom.

Remarks 1. S^2 defined in (*) is also an unbiased estimate of σ^2, since

$$E(S^2) = \frac{\Sigma \nu_i E(S_i^2)}{\Sigma \nu_i} = \frac{(\Sigma \nu_i)\sigma^2}{\Sigma \nu_i} = \sigma^2$$

2. Let S_i^2 and S_j^2 ; $i \neq j$, $1 \leq (i, j) \leq k$ be the smallest and the largest values of the estimates respectively. If on the basis of F-test (*c.f.* Chapter 14), these do not differ significantly, then all the estimates S_t^2 which lie between S_i^2 and S_j^2 won't differ significantly either and consequently all the estimates can be reasonably regarded as homogeneous, coming from the same population. In this case, therefore, there is no need to apply Bartlett's test.

13·11. χ^2-Test for Pooling the Probabilities from Independent Tests to give a Single Test of Significance (P_χ-Test). See Examples 13·1 and 13·2 for detailed discussion.

EXCERCISE 13 (d)

1. Define χ^2-statistic. What are its uses ? What is Fisher's z-transformation for correlation coefficient and what are its properties ? How is it used to combine the correlation coefficients between two random variables computed independently from different sources.

2. Explain the use of the chi-square statistic for testing the homogeneity of several independent estimates of population correlation coefficient, clearly stating the underlying assumptions.

3. (*a*) The correlation coefficients between wing length and tongue length were estimated from 2 samples each of size 44 to be 0·731 and 0·690. Test whether the correlation coefficients are significantly different or not. If not, obtain the best estimate of the common correlation coefficient.

(*b*) Test for equality of the correlation co-efficients between the scores in two halves of a psychological test applied to different groups of sizes 30, 20 and 25 if the corresponding sample values are 0·63, 0·48, 0·71, respectively.

4. (*a*) Independent samples of 21, 30, 39, 26 and 35 pairs of values yielded correlation coefficients 0·39, 0·61, 0·43, 0·54 and 0·48, respectively. Can these estimates be regarded as homogeneous ? If so, find an estimate of the correlation coefficient in the population.

(*b*) Test whether the following set of correlation coefficients between stature and sitting heights obtained for persons from 8 districts can be regarded as homogeneous.

Sample size	130	60	338	78	125	299	170	139
Corr. coefficient :	0·718	0·961	0·825	0·685	0·700	0·548	0·793	0·687

(*c*) The correlation coefficients between fibre weight and staple length in six cotton crosses were estimated as :

$$-0{\cdot}129,\ 0{\cdot}1138,\ -0{\cdot}2780,\ 0{\cdot}0033,\ 0{\cdot}2331 \text{ and } 0{\cdot}0550$$

based on samples of sizes 73, 81, 67, 83, 71, 57 respectively. Test the homogeneity of r_j's and obtain their best estimate.

13·12. Non-central χ^2-distribution. The χ^2-distribution defined as the sum of the squares of independent standard normal variates is often referred to as the *central χ^2-distribution*. The distribution of the sum of the squares of independent normal variates each having unit variance but with possibly non-zero means is known as *non-central chi-square distribution*. Thus if X_i, $(i = 1, 2, \ldots, n)$ are independent $N(\mu_i, 1)$, r.v.'s then

$$\chi'^2 = \sum_{i=1}^{n} X_i^2, \qquad \ldots(13{\cdot}24)$$

has the non-central χ^2 distribution with n d.f. Intuitively, this distribution would seem to depend upon the n parameters $\mu_1, \mu_2, \ldots, \mu_n$ but it will be seen that it depends on these parameters only through the *non-centrality parameter*

$$\lambda = \frac{1}{2}(\mu_1^2 + \mu_2^2 + \ldots + \mu_n^2) \qquad \ldots(13\cdot24a)$$

and we write, $\chi'^2 \sim \chi'^2(n, \lambda)$.

13·12·1. Non-central χ^2-distribution with Non-centrality Parameter λ. *The p.d.f. is given by*

$$f_{\chi_n'^2}(\lambda) = \sum_{j=0}^{\infty}\left[\frac{e^{-\lambda}\lambda^j}{j\,!} \quad p(\chi^2_{n+2j})\right] \qquad \ldots(13\cdot25)$$

where $p(\chi^2_{n+2j})$ is the p.d.f. of (central) χ^2-variate with $n + 2j$ d.f.

Thus $f_{\chi_n'^2}(\lambda)$ is the mixture of central χ^2-distributions with d.f. n, $n + 2$, $n + 4, \ldots$, the corresponding weights being the successive terms of the Poisson distribution with parameter λ.

Derivation of p.d.f. of χ'^2. We shall obtain the p.d.f. of non-central χ^2-distribution through moment generating function (m.g.f.), by using the uniqueness theorem of m.g.f.

13·12·2. Moment Generating Function of Non-central χ^2-Distribution. If $X \sim N(\mu, 1)$ then

$$M_{X^2}(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx^2} \cdot e^{-(x-\mu)^2/2}\,dx$$

$$\exp\left[tx^2 - \frac{1}{2}(x-\mu)^2\right] = \exp\left[-\left\{\left(\frac{1}{2} - t\right)x^2 - \mu x + \frac{\mu^2}{2}\right\}\right]$$

$$= \exp\left[-\left(\frac{1-2t}{2}\right)\left\{x^2 - \frac{2\mu x}{1-2t} + \frac{\mu^2}{1-2t}\right\}\right]$$

$$= \exp\left[-\left(\frac{1-2t}{2}\right)\left\{\left(x - \frac{\mu}{1-2t}\right)^2 + \frac{\mu^2}{1-2t} - \frac{\mu^2}{(1-2t)^2}\right\}\right]$$

$$= \exp\left(\frac{t\mu^2}{1-2t}\right)\exp\left[-\left(\frac{1-2t}{2}\right)\left(x - \frac{\mu}{1-2t}\right)^2\right]$$

$$\therefore\ M_{X^2}(t) = \exp\left(\frac{t\mu^2}{1-2t}\right)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-\left(\frac{1-2t}{2}\right)\left(x - \frac{\mu}{1-2t}\right)^2\right]dx$$

$$= \exp\left(\frac{t\mu^2}{1-2t}\right)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}u^2\right)\frac{du}{(1-2t)^{1/2}}$$

$$= (1-2t)^{-1/2}\exp\left(\frac{t\mu^2}{1-2t}\right);\ 1 - 2t > 0 \Rightarrow t < \frac{1}{2} \qquad \ldots(*)$$

If X_i ($i = 1, 2, \ldots, n$), are independent $N(\mu_i, 1)$ then the m.g.f. of the non-central χ^2-variate $\chi'^2 = \sum_{i=1}^{n} X_i^2$ is given by

$$M_{\chi'^2}(t) = M_{\sum_{i=1}^{n} X_i^2}(t) = \prod_{i=1}^{n} M_{X_i^2}(t) \qquad \text{(since } X_i^2\text{,s are independent)}$$

$$= \prod_{i=1}^{n} \left[(1-2t)^{-1/2} \exp\left(\frac{t\mu_i^2}{1-2t}\right)\right] \qquad \text{[From (*)]}$$

$$= (1-2t)^{-n/2} \exp\left[\frac{t}{(1-2t)} \sum_{i=1}^{n} \mu_i^2\right]$$

$$= (1-2t)^{-n/2}. \exp\left[2\lambda t / (1-2t)\right],\ t < \tfrac{1}{2} \qquad \text{...(13·26)}$$

where $\lambda = \frac{1}{2} \sum_{i=1}^{n} \mu_i^2$, is the non-centrality parameter.

(13·26) can be re-written as

$$M_{\chi'^2}(t) = (1-2t)^{-n/2} \exp. \left[\lambda\left(-1+\frac{1}{1-2t}\right)\right]$$

$$= (1-2t)^{-n/2}\, e^{-\lambda} \exp\left(\frac{\lambda}{1-2t}\right)$$

$$= (1-2t)^{-n/2}\, e^{-\lambda} \sum_{r=0}^{\infty} \left(\frac{\lambda}{1-2t}\right)^r \times \frac{1}{r!}$$

$$= \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} (1-2t)^{-(r+n/2)}\ ;\ t < \tfrac{1}{2} \qquad \text{...(13·26a)}$$

Thus the m.g.f. of a non-central χ^2 distribution is seen to be a *convex-combination* of χ^2 m.g.f.'s with d.f. n, $n+2$, $n+4$, ... The coefficients appearing in the convex combination are merely the Poisson probabilities.

Hence by the uniqueness theorem of m.g.f.'s the p.d.f. of non-central χ^2-distribution with n d.f. and with non-centrality parameter λ is given by

$$f(\chi'^2) = \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} \times p(\chi^2_{n+2r}),$$

where $$p(\chi^2_{n+2r}) = \frac{1}{2^{(n+2r)/2}\,\Gamma\left(\frac{n+2r}{2}\right)} e^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{n}{2}+r-1};\ 0 \le \chi^2 < \infty$$

is the p.d.f. of central χ^2-distribution with $(n+2r)$ d.f.

Remarks 1. We can also write the m.g.f. of non-central χ^2 distribution with non-centrality parameter λ as

$$E\left[((1-2t))^{-\frac{n}{2}-Y}\right],$$

where Y is a Poisson variate with parameter λ.

2. If we take $\lambda = 0 \Rightarrow \mu_i = 0\ \forall\ i = 1, 2, ..., n$, the m.g.f. of the non-central χ^2 distribution reduces to the m.g.f. of central χ^2 distribution; *viz.*, $(1-2t)^{-n/2}$.

3. Taking $\lambda = 0$ in the p.d.f. of non-central χ^2-variate, *i.e.*, in (13·25), we get

$$f(\chi'^2) = p(\chi_n^2) = \frac{1}{2^{n/2}\,\Gamma(n/2)}\, e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}, \quad 0 \le \chi^2 < \infty$$

[$\because$ we get contribution only when $r = 0$, the other terms vanish when $\lambda = 0$]; which is p.d.f. of central χ^2-distribution with n d.f.

13·12·3. Additive or Re-productive Property of Non-central Chi-Square Distribution. *If Y_i, $(i = 1, 2, \ldots, k)$, are independent non-central χ^2-variates with n_i d.f. and non-centrality element λ_i, then $\sum_{i=1}^{k} Y_i$ is also a non-central χ^2-variate with $\sum_{i=1}^{k} n_i$ d.f. and non-centrality element $\lambda = \sum_{i=1}^{k} \lambda_i$.*

Proof. We have from (13·26),

$$M_{Y_i}(t) = (1 - 2t)^{-n_i/2} \exp\left[2t\,\lambda_i / (1 - 2t)\right], \quad (i = 1, 2, \ldots, k)$$

$$\therefore \quad M_{\sum Y_i}(t) = \prod_{i=1}^{k} M_{Y_i}(t) = (1 - 2t)^{-\sum_i n_i/2} \exp\left[2t \sum_i \lambda_i / (1 - 2t)\right],$$

which is the m.g.f. of a non-central χ^2-variate with $\sum_i n_i$ d.f. and non-centrality parameter $\lambda = \sum_i \lambda_i$. Hence by uniqueness theorem of m.g.f.'s

$$\sum_{i=1}^{k} Y_i \sim \chi'^2_{\sum n_i} \left(\sum_i \lambda_i\right)$$

13·12·4. Cumulants of Non-central Chi-square Distribution. Cumulant generating function is given by

$$K_{\chi'^2}(t) = \log M_{\chi'^2}(t) = -\frac{n}{2} \log\,(1 - 2t) + 2t\lambda\,(1 - 2t)^{-1}$$

$$= \frac{n}{2}\left[2t + \frac{(2t)^2}{2} + \ldots + \frac{(2t)^r}{r} + \ldots\right] + 2\lambda t\left[1 + 2t + (2t)^2 + \ldots + (2t)^r + \ldots\right]$$

the expansion being valid for $t < 1/2$.

$$\therefore K_{\chi'^2}(t) = (n + 2\lambda)t + (n + 4\lambda)\, t^2 + \ldots + \left(\frac{2^{r-1}}{r} \cdot n + 2\lambda\, 2^{r-1}\right) t^r + \ldots$$

$$\Rightarrow \quad \kappa_r = \text{Coefficient of } \frac{t^r}{r\,!} \text{ in } K_{\chi'^2}(t) = r\,!\left(\frac{n}{r} + 2\lambda\right) 2^{r-1}$$

$$= 2^{r-1}(r - 1)\,!\,(n + 2\lambda r) \qquad \ldots(13·27)$$

$$\therefore \quad \kappa_{r-1} = 2^{r-2}\,(r - 2)\,!\,[n + 2\lambda\,(r - 1)]$$

$$\therefore \quad \frac{d}{d\lambda}\{\kappa_{r-1}\} = 2^{r-2}\,(r - 2)\,!\,2(r - 1) = 2^{r-1}\,(r - 1)\,!$$

$$= \frac{\kappa_r}{(n + 2\lambda r)} \qquad \text{[From (13·27)]}$$

$$\Rightarrow \quad \kappa_r = (n + 2\lambda r)\,\frac{d}{d\lambda}(\kappa_{r-1}) \qquad \ldots(13·28)$$

Exact Sampling Distributions

(CONTINUED)

(t, F AND Z DISTRIBUTIONS)

14·1. Introduction. The entire large sample theory was based on the application of "*Normal Test*" (*c.f.* § 12·9). However, if the sample size n is small, the distribution of the various statistics, *e.g.*, $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ or $Z = (X - nP)/\sqrt{nPQ}$ etc., are far from normality and as such '*normal test*' cannot be applied if n is small. In such cases exact sample tests, pioneered by W.S. Gosset (1908) who wrote under the pen name of Student, and later on developed and extended by Prof. R.A. Fisher (1926), are used. In the following sections we shall discuss

(*i*) *t-test*, (*ii*) *F-test, and* (*iii*) *Fisher's z-transformation.*

The exact sample tests can, however, be applied to large samples also though the converse is not true. In all the exact sample tests, the basic assumption is that "*The population(s) from which sample(s) are drawn is (are) normal, i.e., the parent population(s) is (are) normally distributed.*"

14·2. Student's 't'. *Definition.* Let x_i, (i = 1, 2, ..., n) be a random sample of size n from a normal population with mean μ and variance σ^2. Then Student's t is defined by the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \text{...(14·1)}$$

where $\bar{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$, is the sample mean and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad \text{...(14·1a)}$$

is an unbiased estimate of the population variance σ^2, and it follows Student's t-distribution with $\nu = (n - 1)$ *d.f.* with probability density function,

$$f(t) = \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu + 1)/2}} ; -\infty < t < \infty \qquad \text{...(14·2)}$$

Remarks 1. A statistic t following Student's t-distribution with n d.f. will be abbreviated as $t \sim t_n$.

2. If we take $\nu = 1$ in (14·2), we get

$$f(t) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{1}{(1 + t^2)},$$

$$= \frac{1}{\pi} \cdot \frac{1}{(1 + t^2)}\ ; -\infty < t < \infty \qquad [\because \Gamma(\tfrac{1}{2}) = \sqrt{\pi}]$$

which is the p.d.f. of standard Cauchy distribution. Hence, when $\nu = 1$ Student's t distribution reduces to Cauchy distribution.

14·2·1. Derivation of Student's t-distribution. The expression (14·1) can be re-written as

$$t^2 = \frac{n(\bar{x} - \mu)^2}{S^2} = \frac{n(\bar{x} - \mu)^2}{ns^2/(n - 1)} \qquad [\because ns^2 = (n - 1)\, S^2]$$

$$\Rightarrow \quad \frac{t^2}{(n - 1)} = \frac{(\bar{x} - \mu)^2}{\sigma^2/n} \ \frac{1}{ns^2/\sigma^2} = \frac{(\bar{x} - \mu)^2/(\sigma^2/n)}{ns^2/\sigma^2}$$

Since x_i, $(i = 1, 2, \ldots, n)$ is a random sample from the normal population with mean μ and variance σ^2,

$$\bar{x} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Hence $\dfrac{(\bar{x} - \mu)^2}{\sigma^2/n}$, being the square of a standard normal variate is a chi-square variate with 1 d.f.

Also $\dfrac{ns^2}{\sigma^2}$ is a χ^2-variate with $(n - 1)$ d.f. (*c.f. Theorem 13·5).*

Further since $\bar{x}$ and s^2 are independently distributed (*c.f. Theorem* 13·5), $\dfrac{t^2}{n - 1}$, being the ratio of two independent χ^2-variates with 1 and $(n - 1)$ d.f. respectively, is a $\beta_2\left(\frac{1}{2}, \frac{n - 1}{2}\right)$ variate and its distribution is given by :

$$dF(t) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{(t^2/\nu)^{\frac{1}{2} - 1}}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu + 1)/2}}\, d(t^2/\nu),\ 0 \le t^2 < \infty$$

[where $\nu = (n - 1)$]

$$= \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu + 1)/2}}\, dt; -\infty < t < \infty$$

the factor 2 disappearing since the integral from $-\infty$ to ∞ must be unity. This is the required probability function as given in (14·2) of Student's t-distribution with $\nu = (n - 1)$ d.f.

Remarks on Student's 't'. 1. *Importance of Student's t-distribution in Statistics.* W.S. Gosset, who wrote under pseudonym (pen-name) of Student defined his t in a slightly different way, *viz.*, $t = (\bar{x} - \mu)/s$ and investigated its sampling distribution, somewhat empirically, in a paper entitled *'The probable error of the mean'*, published in 1908. Prof. R.A. Fisher, later on defined his own 't' and gave a rigorous proof for its sampling distribution in 1926. The salient feature of 't' is that both the statistic and its sampling distribution are functionally independent of σ, the population standard deviation.

The discovery of 't' is regarded as a landmark in the history of statistical inference because of the following reason. Before Student gave his 't' it was customary to replace σ^2 in $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, by its unbiased estimate S^2 to give $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$ and then normal test was applied even for small samples. It has been found that although the distribution of t is asymptotically normal for large n (*c.f.* § 14·2·5), it is far from normality for small samples. The Student's t ushered in an era of exact sample distributions (and tests) and since its discovery many important contributions have been made towards the development and extension of small (exact) sample theory.

2. *Confidence or Fiducial Limits for μ.* If $t_{0.05}$ is the tabulated value of t for $\nu = (n - 1)$ d.f. at 5% level of significance, *i.e.*,

$$P(|t| > t_{0.05}) = 0{\cdot}05 \Rightarrow P(|t| \le t_{0.05}) = 0{\cdot}95,$$

the 95% confidence limits for μ are given by :

$$|t| \le t_{0.05}, \quad i.e., \quad \left|\frac{\bar{x} - \mu}{S/\sqrt{n}}\right| \le t_{0.05}$$

$$\Rightarrow \qquad \bar{x} - t_{0.05} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{x} + t_{0.05} \frac{S}{\sqrt{n}}$$

Thus, 95% confidence limits for μ are :

$$\bar{x} \pm t_{0.05} \cdot \frac{S}{\sqrt{n}} \qquad \text{...[14·2(a)]}$$

Similarly, 99% confidence limits for μ are :

$$\bar{x} \pm t_{0.01} \frac{S}{\sqrt{n}} \qquad \text{...[14·2(b)]}$$

where $t_{0.01}$ is the tabulated value of t for $\nu = (n - 1)$ d.f. at 1% level of significance.

14·2·2. Fisher's 't' (*Definition*). It is the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom. If ξ is a $N(0, 1)$ and χ^2 is an independent chi-square variate with n *d.f.*, then Fisher's t is given by

$$t = \xi \bigg/ \sqrt{\frac{\chi^2}{n}} \qquad \text{...(14·3)}$$

and it follows student's 't' distribution with n degrees of freedom.

14·2·3. Distribution of Fisher's 't'. Since ξ and χ^2 are independent, their joint probability differential is given by

$$dF(\xi, \chi^2) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-\xi^2/2) \frac{\exp(-\chi^2/2)(\chi^2)^{\frac{n}{2}-1}}{2^{n/2}\,\Gamma(n/2)} d\xi\, d\chi^2$$

Let us transform to new variates t and u by the substitution

$$t = \frac{\xi}{\sqrt{\chi^2/n}} \text{ and } u = \chi^2 \quad \Rightarrow \quad \xi = t\sqrt{u/n} \text{ and } \chi^2 = u$$

Jacobian of transformation J is given by

$$J = \frac{\partial(\xi, \chi^2)}{\partial(t, u)} = \begin{vmatrix} \sqrt{u/n} & t/(2\sqrt{un}) \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{u}{n}}$$

The joint distribution of t and u becomes

$$dG(t,u) = \frac{1}{\sqrt{2\pi}\, 2^{n/2}\,\Gamma(n/2)\sqrt{n}} \exp\left\{-\frac{u}{2}\left(1 + \frac{t^2}{n}\right)\right\} u^{\frac{n}{2}-\frac{1}{2}} du\, dt\,;$$

Integrating w.r.t. 'u' over the range 0 to ∞, the marginal distribution of t becomes

$$dG_1(t) = \frac{1}{\sqrt{2\pi}\, 2^{n/2}\,\Gamma(n/2)\sqrt{n}} \left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1 + \frac{t^2}{n}\right)\right\} u^{(n-1)/2} du\right] dt$$

$$= \frac{1}{\sqrt{2\pi}\, 2^{n/2}\,\Gamma(n/2)\sqrt{n}} \frac{\Gamma[(n+1)/2]}{\left[\frac{1}{2}\left(1 + \frac{t^2}{n}\right)\right]^{(n+1)/2}} dt$$

$$\therefore\; dG_1(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n}\,\Gamma(n/2)\Gamma(\frac{1}{2})} \cdot \frac{1}{\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}} dt, \; -\infty < t < \infty$$

$$= \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}} dt, \; -\infty < t < \infty$$

which is same as the probability function of Student's t-distribution with n d.f.

Remarks 1. In Fisher's 't' the d.f. is the same as the d.f. of chi-square variate.

2. Student's 't' may be regarded as a particular case of Fisher's 't' as explained below.

$$\text{Since } \bar{x} \sim N(\mu, \sigma^2/n), \qquad \xi = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad \ldots(*)$$

and

$$\chi^2 = \frac{ns^2}{\sigma^2} = \sum_{i=1}^{n} (x_i - \bar{x})^2/\sigma^2 \qquad \ldots(**)$$

is independently distributed as chi-square variate with $(n-1)$ d.f. Hence Fisher's t is given by

$$t = \frac{\xi}{\sqrt{\chi^2/(n-1)}} = \frac{\sqrt{n}\,(\bar{x}-\mu)}{\sigma} \cdot \frac{\sigma}{\sqrt{\Sigma(x_i-\bar{x})^2/(n-1)}}$$

$$= \frac{\sqrt{n}\,(\bar{x}-\mu)}{S} = \frac{\bar{x}-\mu}{S/\sqrt{n}} \qquad \ldots(***)$$

and it follows Student's t-distribution with $(n-1)$ d.f. (*c.f.* Remark 1 above.)

Now, (***) is same as Student's 't' defined in (14·1). Hence Student's 't' is a particular case of Fisher's 't'.

14·2·4. Constants of t-distribution. Since $f(t)$ is symmetrical about the line $t = 0$, all the moments of odd order about origin vanish, *i.e.*,

$$\mu'_{2r+1} \text{ (about origin)} = 0 \; ; \; r = 0, 1, 2, \ldots$$

In particular,

$$\mu_1' \text{ (about origin)} = 0 = \text{Mean}$$

Hence central moments coincide with moments about origin.

$$\therefore \qquad \mu_{2r+1} = 0, \quad (r = 1, 2, \ldots) \qquad \ldots(14\cdot4)$$

The moments of even order are given by

$$\mu_{2r} = \mu'_{2r} \text{ (about origin)}$$

$$= \int_{-\infty}^{\infty} t^{2r} f(t)\,dt = 2\int_0^{\infty} t^{2r} f(t)\,dt$$

$$= 2 . \frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)\sqrt{n}} \int_0^{\infty} \frac{t^{2r}}{\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}}\,dt$$

This integral is absolutely convergent if $2r < n$.

Put $1 + \frac{t^2}{n} = \frac{1}{y} \;\Rightarrow\; t^2 = n(1-y)/y$ *i.e.*, $2t\,dt = -\frac{n}{y^2}\,dy$

When $t = 0$, $y = 1$ and when $t = \infty$, $y = 0$. Therefore,

$$\mu_{2r} = \frac{2}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_1^0 \frac{t^{2r}}{(1/y)^{(n+1)/2}} \cdot \frac{-n}{2ty^2}\,dy$$

$$= \frac{n}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 (t^2)^{(2r-1)/2}\, y^{\{(n+1)/2\}-2}\,dy$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 \left[n\left(\frac{1-y}{y}\right)\right]^{r-\frac{1}{2}} y^{\{(n+1)/2\}-2}\,dy$$

$$= \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 y^{\frac{n}{2} - r - 1} (1 - y)^{r - \frac{1}{2}} dy$$

$$= \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} . B\left(\frac{n}{2} - r\ ,\ r + \frac{1}{2}\right), n > 2r. \qquad ...[14·4(a)]$$

$$= n^r \frac{\Gamma[(n/2) - r]\, \Gamma(r + \frac{1}{2})}{\Gamma(\frac{1}{2})\, \Gamma(n/2)}$$

$$= n^r \frac{(r - \frac{1}{2})(r - \frac{3}{2}) \ldots \frac{3}{2}\, \frac{1}{2}\, \Gamma(\frac{1}{2})\, \Gamma[(n/2) - r]}{\Gamma(\frac{1}{2})\, [(n/2) - 1][(n/2) - 2]\ldots[(n/2) - r]\Gamma[(n/2) - r]}$$

$$= n^r \frac{(2r - 1)(2r - 3)\ldots 3\cdot 1}{(n - 2)(n - 4)\ldots(n - 2r)}, \frac{n}{2} > r \qquad ...[14·4\ (b)]$$

In particular

$$\mu_2 = n \frac{1}{(n - 2)} = \frac{n}{n - 2}, [n > 2] \qquad ...[14·4(c)]$$

and

$$\mu_4 = n^2 \frac{3\cdot 1}{(n - 2)(n - 4)} = \frac{3n^2}{(n - 2)(n - 4)}, [n > 4] \qquad ...[14·4(d)]$$

Hence

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = 3\left(\frac{n - 2}{n - 4}\right)$$

Remarks 1. As $n \to \infty$, $\beta_1 = 0$ and

$$\beta_2 = \lim_{n \to \infty} 3\left(\frac{n - 2}{n - 4}\right) = 3 \lim_{n \to \infty} \left[\frac{1 - (2/n)}{1 - (4/n)}\right] = 3 \qquad ...[14·4(e)]$$

2. Changing r to $(r - 1)$ in [14·4(b)], dividing and simplifying, we shall get the recurrence relation for the moments as

$$\frac{\mu_{2r}}{\mu_{2r - 2}} = \frac{n(2r - 1)}{(n - 2r)}, \frac{n}{2} > r \qquad ...[14·4(c)]$$

3. Moment Generating Function of t-distribution. From [14·4(b)] we observe that if $t \sim t_n$, then all the moments of order $2r < n$ exist but the moments of order $2r \geq n$ do not exist. Hence the m.g.f. of t-distribution does not exist.

Example 14·1. *Express the constants y_0, a and m of the distribution :*

$$dF(x) = y_0 \left[1 - \frac{x^2}{a^2}\right]^m dx, \ -a \leq x \leq a \qquad ...(*)$$

in terms of its μ_2 and β_2.

Show that if x is related to a variable t by the equation

$$x = \frac{at}{\{2(m + 1) + t^2\}^{1/2}}, \qquad ...(**)$$

then t has Student's distribution with $2(m + 1)$ degrees of freedom. Use the transformation to calculate the probability that $t \geq 2$ when the degrees of freedom are 2 and also when 4. **(Madras Univ. M.Sc., 1991)**

Solution. First of all we shall determine the constant from the consideration that total probability is unity.

$$\therefore \qquad y_0 \int_{-a}^{a} \left(1 - \frac{x^2}{a^2}\right)^m dx = 1$$

$$\Rightarrow \qquad 2y_0 \int_{0}^{a} \left(1 - \frac{x^2}{a^2}\right)^m dx = 1$$

($\because$ Integrand is an even function of x)

$$\Rightarrow \qquad 2y_0 \int_0^{\pi/2} \cos^{2m}\theta \,.\, a\cos\theta \, d\theta = 1 \qquad (x = a \sin \theta)$$

$$\Rightarrow \qquad 2ay_0 \int_0^{\pi/2} \cos^{2m+1}\theta \, d\theta = 1$$

But we have the Beta integral,

$$2\int_0^{\pi/2} \sin^p\theta \cos^q\theta \, d\theta = B\left(\frac{p+1}{2}, \frac{q+1}{2}\right) \qquad \ldots(1)$$

$$\therefore \qquad ay_0 . 2\int_0^{\pi/2} \cos^{2m+1}\theta \sin^0\theta \, d\theta = 1$$

$$\Rightarrow \qquad ay_0 \, B(m+1, \tfrac{1}{2}) = 1 \qquad \text{[Using (1)]}$$

$$\Rightarrow \qquad y_0 = \frac{1}{a\, B(m+1, \frac{1}{2})} \qquad \ldots(2)$$

Since the given probability function is symmetrical about the line $x = 0$, we have as in § 14·2·4.

$$\mu_{2r+1} = \mu_{2r+1}' = 0; \; r = 0, 1, 2, \ldots \qquad [\because \text{Mean} = \text{Origin}]$$

The moments of even order are given by

$$\mu_{2r} = \mu_{2r}' \text{ (about origin)}$$

$$= \int_{-a}^{a} x^{2r} f(x)\, dx = y_0 \int_{-a}^{a} x^{2r} \left(1 - \frac{x^2}{a^2}\right)^m dx$$

$$= 2y_0 \int_0^a x^{2r} \left(1 - \frac{x^2}{a^2}\right)^m dx$$

$$= 2y_0 \int_0^{\pi/2} (a \sin \theta)^{2r} \cos^{2m} \theta.\ a \cos \theta\, d\theta \qquad [x = a \sin \theta]$$

$$= y_0\, a^{2r+1}.\, 2 \int_0^{\pi/2} \sin^{2r} \theta \,.\, \cos^{2m+1} \theta\, d\theta$$

$$= y_0\, a^{2r+1}\, B(r + \tfrac{1}{2},\, m + 1) \qquad \text{[Using (1)]}$$

$$= a^{2r} \frac{B(r + \frac{1}{2},\, m + 1)}{B(m + 1,\, \frac{1}{2})} = a^{2r} \,.\, \frac{\Gamma\left(r + \frac{1}{2}\right)\Gamma\left(m + \frac{3}{2}\right)}{\Gamma\left(m + r + \frac{3}{2}\right)\Gamma\left(\frac{1}{2}\right)} \qquad \ldots(***)$$

In particular, $\mu_2 = a^2 \,.\, \dfrac{\Gamma\{m + (3/2)\} \,.\, \frac{1}{2}\Gamma(1/2)}{\{m + (3/2)\}\ \Gamma\{m + (3/2)\}\ \Gamma(1/2)} = \dfrac{a^2}{2m + 3}$

$\therefore \qquad a^2 = (2m + 3)\mu_2 \qquad \ldots(3)$

Also $\qquad \mu_4 = a^4 \dfrac{\Gamma(5/2)}{\Gamma\{m + (7/2)\}} \times \dfrac{\Gamma\{m + (3/2)\}}{\Gamma(1/2)}$

$$= \frac{3a^4}{(2m + 5)\ (2m + 3)} \qquad \textit{(On simplification)}$$

$\therefore \qquad \beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3(2m + 3)}{(2m + 5)}$

$\Rightarrow \qquad m = \dfrac{9 - 5\beta_2}{2(\beta_2 - 3)} \qquad$ *(On simplification)* ... (4)

Equations (2), (3) and (4) express the constants y_0, a and m in terms of μ_2 and β_2.

$$x = \frac{at}{[2(m + 1) + t^2]^{1/2}} \quad \Rightarrow \quad \frac{x^2}{a^2} = \frac{t^2}{2(m + 1) + t^2}$$

i.e., $\qquad 1 - \dfrac{x^2}{a^2} = \dfrac{2(m + 1)}{2(m + 1) + t^2} = \left(1 + \dfrac{t^2}{n}\right)^{-1}, \ (n = 2m + 2)$

Also $\qquad dx = a\left[\dfrac{dt}{(n + t^2)^{1/2}} - t \,.\, \dfrac{1}{2}\, \dfrac{2t\, dt}{(n + t^2)^{3/2}}\right]$

$$= a \frac{1}{(n + t^2)^{1/2}} \left[1 - \frac{t^2}{n + t^2}\right] dt$$

$$= \frac{an}{(n + t^2)^{3/2}}\, dt = \frac{a}{\sqrt{n}} \,.\, \frac{1}{[1 + (t^2/n)]^{3/2}}\, dt$$

Hence the p.d.f. of X transforms to

$$dF(t) = y_0 \frac{1}{\left[1 + \dfrac{t^2}{n}\right]^m} \,.\, \frac{a}{\sqrt{n}} \frac{dt}{\left[1 + \dfrac{t^2}{n}\right]^{3/2}}$$

$$= \frac{1}{a\,B\left(m+1, \frac{1}{2}\right)} \cdot \frac{a}{\sqrt{n}} \frac{dt}{\left[1+\frac{t^2}{n}\right]^{m+(3/2)}}$$

$$= \frac{1}{\sqrt{n}\,B\left(\frac{n}{2}, \frac{1}{2}\right)} \cdot \frac{dt}{\left[1+\frac{t^2}{n}\right]^{(n+1)/2}}, \quad -\infty < t < \infty \quad \ldots(5)$$

which is the probability differential of Student's t-distribution with $n = 2(m+1)$ d.f. Hence the result.

For 2 d.f. *i.e.*, $n = 2$, we get $2(m+1) = 2 \Rightarrow m = 0$. Hence from (**), we get (for $m = 0$),

$$x = \frac{at}{(2+t^2)^{1/2}} \quad \Rightarrow \quad x = \frac{\sqrt{2}}{\sqrt{3}} a, \text{ when } t = 2.$$

$$\therefore \qquad P(t \geq 2) = P\left(X \geq \sqrt{(2/3)}\,a\right) = \int_{a\sqrt{(2/3)}}^{a} dF(x)$$

$$= \int_{a\sqrt{(2/3)}}^{a} \frac{1}{a\,B(1, \frac{1}{2})} dx \qquad \text{[From (*), since } m = 0\text{]}$$

$$= \frac{1}{2a}\left(a - \frac{\sqrt{2}}{\sqrt{3}} a\right) = \frac{\sqrt{3}-\sqrt{2}}{2\sqrt{3}}$$

$$\left[\because B(1, \tfrac{1}{2}) = \frac{\Gamma 1\,\Gamma(1/2)}{\Gamma(3/2)} = \frac{\Gamma(1/2)}{(1/2)\,\Gamma(1/2)} = 2\right]$$

For 4 d.f., *i.e.*, $n = 4$, we get $m = 1$. Proceeding exactly similarly we shall obtain

$$P(t \geq 2) = \frac{1}{2} - \frac{5\sqrt{2}}{16}$$

EXERCISE 14(a)

1. (*a*) Given that

   (*i*) u is normally distributed with zero mean and unit variance,

   (*ii*) v^2 has a chi-square distribution with n degrees of freedom, and

   (*iii*) u and v are independently distributed,

find the distribution of the variable

$$t = \frac{u\sqrt{n}}{v}$$

(*b*) Find the variance of the t distribution with n degrees of freedom, ($n > 2$).

(*c*) If the variable t has Student's t distribution with 2 degrees of freedom, prove that

$$P(t \geq 2) = \frac{3 - \sqrt{6}}{6}$$

[*Shivaji Univ. B.Sc., 1990*]

2. (*a*) State, (without proof), the sampling distribution of Student's t. Who discovered it?

(*b*) 'Discovery of Student's t is regarded as a landmark in the history of statistical inference'. Elucidate.

(*c*) Let t be distributed as Student's t-distribution with 2 d.f. Find the probability $P(-\sqrt{2} \leq t \leq \sqrt{2})$.

3. (*a*) Show that

$$E(T^r) = \begin{cases} \dfrac{k^{r/2}\,\Gamma\left(\dfrac{1+r}{2}\right) . \Gamma\left(\dfrac{k-r}{2}\right)}{\Gamma(1/2) . \Gamma(k/2)}, & \text{if } r \text{ is even for } -1 < r < k \\ 0, \text{ if } r \text{ is odd} \end{cases}$$

where T has Student's t-distribution with k degrees of freedom.

(*b*) For the t-distribution with n d.f., establish the recurrence relation

$$\mu_{2r} = \frac{n\,(2r-1)}{(n-2r)} . \mu_{2r-2},\ n > 2r$$

[*Poona Univ. B.Sc., 1990; Delhi Univ. B.Sc. (Stat. Hons.), 1992*]

(*c*) For how many d.f. does (*i*) χ^2-distribution reduce to negative exponential distribution and (*ii*) t-distribution reduce to Cauchy distribution?

4. Suppose $X_1, X_2, \ldots, X_n$ $(n > 1)$ are independent variates each distributed as $N(0, \sigma^2)$. Find the p.d.f. of

$$W = X_1 \Big/ \left\{\frac{1}{n} \sum_{i=1}^{n} X_i^2\right\}^{1/2}$$

Why does not W follow the t-distribution?

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

5. Let $x_1, x_2, \ldots, x_n$ be independent observations from a normal universe with mean μ and variance σ^2 and let $\bar{x}$ and s^2 be the sample mean and sum of the squares of the deviations from the mean respectively. Let x' be one more observation independent of previous ones. Show that

$$\frac{x' - \bar{x}}{s}\left[\frac{n(n-1)}{n+1}\right]^{1/2}$$

has a Student t-distribution with $(n-1)$ degrees of freedom.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

6. (*a*) Let X_1 and X_2 be two independent normal variates with the same normal distribution $N(\mu, \sigma^2)$. Obtain the distribution of

$$Y = \frac{X_1 + X_2 - 2\mu}{\sqrt{|X_1 - X_2|^2}}$$

Ans. Standard Cauchy distribution.

(b) If X is t-distributed with k degrees of freedom, show that

$$\frac{1}{1+(X^2/k)},$$

has a beta distribution.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1988*]

7. Define Student's t-statistic and state its probability density function.

If x_i $(i = 1, 2, \ldots, n)$, is a random sample of n independent observations from a normal population with mean μ and variance σ^2, show that

$$U = \frac{(\bar{x} - \mu)\sqrt{n(n-1)}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

conforms to Student's t-variate. If x is an additional observation drawn independently from the same normal population, show that

$$W = \frac{(x - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \times \sqrt{\frac{n(n-1)}{n+1}}$$

also conforms to Student's t-variate.

8. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, and $\bar{X}$ and S^2, respectively, be the sample mean and sample variance. Let $X_{n+1} \sim N(\mu, \sigma^2)$, and assume that $X_1, X_2, \ldots, X_n, X_{n+1}$ are independent. Obtain the sampling distribution of

$$U = \frac{(X_{n+1} - \bar{X})}{S} \cdot \sqrt{\frac{n}{n+1}}; \quad \left[S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

9. If the random variables X_1 and X_2 are independent and follow chi-square distribution with n d.f., show that $\dfrac{\sqrt{n}(X_1 - X_2)}{2\sqrt{X_1 X_2}}$ is distributed as Student's t with n d.f., independently of $X_1 + X_2$.

[*Calcutta Univ. B.Sc. (Hons.), 1992*]

Hint. $p(x_1, x_2) = \dfrac{1}{2^n[\Gamma(n/2)]^2} \cdot e^{-(x_1+x_2)/2}\, x_1^{(n/2)-1}\, x_2^{(n/2)-1};$

$$0 \le x_1 < \infty, \; 0 \le x_2 < \infty$$

Put $\quad u = \dfrac{\sqrt{n}(x_1 - x_2)}{2\sqrt{x_1 x_2}}$ and $v = x_1 + x_2$

$$\Rightarrow \quad x_1 = \frac{v}{2}\left[1 + \frac{1}{\sqrt{\left(1 + \frac{n}{u^2}\right)}}\right], \quad x_2 = \frac{v}{2}\left[1 - \frac{1}{\sqrt{\left(1 + \frac{n}{u^2}\right)}}\right]$$

Jacobian of transformation is $J = \dfrac{\partial(x_1, x_2)}{\partial(u, v)} = \dfrac{v}{2\sqrt{n}\,[1 + u^2/n]^{3/2}}$

The joint p.d.f. of U and V becomes

$$g(u, v) = p(x_1, x_2)\,|J| = \frac{1}{2^{2n-1}\,\Gamma(n/2)\,\Gamma(n/2)\,\sqrt{n}} \cdot \frac{e^{-v/2}\,v^{n-1}}{(1 + u^2/n)^{(n+1)/2}}\,;$$

$$-\infty < u < \infty,\ 0 \le v < \infty$$

Using Legender's duplication formula, *viz.*,

$$\Gamma n = 2^{n-1}\,\Gamma(n/2)\,\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi} \;\Rightarrow\; \Gamma(n/2) = \frac{\Gamma n\,\sqrt{\pi}}{2^{n-1}\,\Gamma\left(\frac{n+1}{2}\right)}, \text{ we get}$$

$$2^{2n-1}\,\Gamma(n/2)\,\Gamma(n/2)\,\sqrt{n} = \frac{2^{2n-1}\cdot\sqrt{n}\,\sqrt{\pi}}{2^{n-1}\,\Gamma\left(\frac{n+1}{2}\right)}\,\Gamma\left(\frac{n}{2}\right)\sqrt{n}$$

$$= 2^{n}\,\sqrt{n}\,\sqrt{n}\,B\left(\tfrac{1}{2}, n/2\right) \qquad [\because \sqrt{\pi} = \Gamma(\tfrac{1}{2})]$$

$$g(u, v) = \left(\frac{1}{2^n\,\Gamma n}\,e^{-v/2}\,v^{n-1}\right)\left[\frac{1}{\sqrt{n}\,B\left(\frac{1}{2}, n/2\right)} \cdot \frac{1}{\left(1 + \dfrac{u^2}{n}\right)^{(n+1)/2}}\right];$$

$$0 < v < \infty,\ -\infty < u < \infty.$$

10. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. If $\bar{X}$ and $\bar{Y}$ denote the corresponding sample means and if

$$(m-1)S_1^2 = \sum_{i=1}^{m}(X_i - \bar{X})^2,\ (n-1)\,S_2^2 = \sum_{j=1}^{n}(Y_j - \bar{Y})^2,$$

obtain the sampling distribution of

$$\frac{a(\bar{X} - \mu_1) + b(\bar{Y} - \mu_2)}{\left[\left\{\dfrac{(m-1)\,S_1^2 + (n-1)\,S_2^2}{(m+n-2)}\right\}\left\{\dfrac{a^2}{m} + \dfrac{b^2}{n}\right\}\right]^{1/2}}$$

where a and b are two fixed real numbers.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

11. If $I_x(p, q)$ represents the incomplete Beta function defined by

$$I_x(p, q) = \frac{1}{B(p, q)}\int_0^x t^{p-1}\,(1-t)^{q-1}\,dt;\ p > 0,\ q > 0,$$

show that the distribution function $F(.)$ of Student's t-distribution is given by

$$F(t) = 1 - \tfrac{1}{2}\,I_x\left(\frac{n}{2}, \frac{1}{2}\right), \text{ where } x = \left(1 + \frac{t^2}{n}\right)^{-1}.$$

[Delhi Univ. M.Sc. (Stat.), 1990; Nagpur Univ. M.Sc. (Stat.), 1991]

Hint. If $f(.)$ is p.d.f. of t-distribution with n d.f. then

$$F(t) = \int_{-\infty}^{t} f(u)\, du = 1 - \int_{t}^{\infty} f(u)\, du$$

$$= 1 - \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{t}^{\infty} \left(1 + \frac{u^2}{n}\right)^{-(n+1)/2} du$$

$$= 1 + \frac{1}{2\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{\left(1 + \frac{t^2}{n}\right)^{-1}}^{0} z^{(n/2)-1} (1-z)^{-1/2}\, dz,$$

where $\left[\frac{1}{z} = 1 + \frac{u^2}{n}\right]$

$$= 1 - \frac{1}{2\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{0}^{x} z^{(n/2)-1} (1-z)^{-1/2}\, dz, \quad \left[x = \left(1 + \frac{t^2}{n}\right)^{-1}\right]$$

$$= 1 - \frac{1}{2} I_x\left(\frac{n}{2}, \frac{1}{2}\right)$$

12. Show that for t-distribution with n d.f., mean deviation about mean is given by

$$\sqrt{n}\, \Gamma\left(\frac{n-1}{2}\right) \Big/ \sqrt{\pi}\, \Gamma(n/2)$$

(*Shivaji Univ. B.Sc. Oct., 1992*]

Hint. $E(t) = 0$.

$$\text{M.D. about mean} = \int_{-\infty}^{\infty} |t| f(t)\, dt$$

$$= \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{-\infty}^{\infty} \frac{|t|\, dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}$$

$$= \frac{2}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{0}^{\infty} \frac{t\, dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{0}^{\infty} \frac{dy}{(1 + y)^{(n+1)/2}}, \quad \left[\left(\frac{t^2}{n} = y\right)\right]$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^\infty \frac{y^{1-1}}{(1+y)^{\frac{n-1}{2}+1}}\, dy$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)}\, B\left(\frac{n-1}{2}, 1\right)$$

13. If $X \sim t_{(n)}$, show that

$$(n - \tfrac{1}{2}) \log\left[1 + \frac{x^2}{n}\right] \sim \chi^2_{(1)},$$

for large n.

You may assume that for large n,

$$\frac{\Gamma\left(\frac{n}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2} + n\right)\sqrt{\frac{1}{2}n}} \simeq \left(1 - \frac{1}{4n}\right)$$

14. If $\bar{X}$ and $\hat{\sigma}^2 = S^2$ be the usual sample mean and sample variance based on a random sample of n observations from $N(\mu, \sigma^2)$, and if $T = (\bar{X} - \mu)\sqrt{n}/S$, prove that

(i) $\text{Var}(T) = (n-1)/(n-3)$

(ii) $\text{Cov}(\bar{X}, T) = \sigma \dfrac{\sqrt{n-1}\ \Gamma[(n-2)/2]}{\sqrt{2n}\ \Gamma[(n-1)/2]}$

(iii) $r(\bar{X}, T) = [\frac{1}{2}(n-3)]^{1/2}\, \Gamma[\frac{1}{2}(n-2)] / \Gamma[\frac{1}{2}(n-1)]$

14·2·5. Limiting Form of t-distribution. *As $n \to \infty$, the p.d.f. of t-distribution with n d.f. viz.,*

$$f(t) = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \to \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \ -\infty < t < \infty$$

Proof. $$\lim_{n\to\infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{\Gamma[(n+1)/2]}{\Gamma(\frac{1}{2})\, \Gamma(n/2)}$$

$$= \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\pi}} \left(\frac{n}{2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2\pi}}.$$

$$\left[\because \Gamma(\tfrac{1}{2}) = \sqrt{\pi} \text{ and } \lim_{n\to\infty} \frac{\Gamma(n+k)}{\Gamma(n)} = n^k, \text{ (c.f. Remark to § 14·5·7)}\right]$$

$$\therefore \lim_{n \to \infty} f(t) = \lim_{n \to \infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \lim_{n \to \infty} \left[\left(1 + \frac{t^2}{n}\right)^n\right]^{-\frac{1}{2}}$$

$$\times \lim_{n \to \infty} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}}$$

$$= \frac{1}{\sqrt{2\pi}} \exp(-t^2/2), \quad -\infty < t < \infty$$

Hence for large d.f. t-distribution tends to standard normal distribution.

14·2·6. Graph of t-distribution. The p.d.f. of t-distribution with n d.f. is

$$f(t) = C \cdot \left[1 + \frac{t^2}{n}\right]^{-(n+1)/2}, \quad -\infty < t < \infty$$

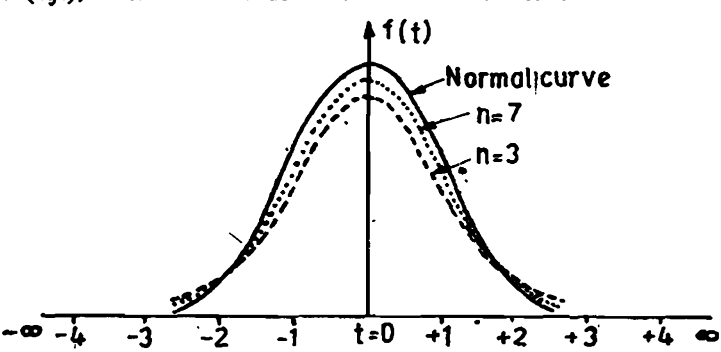
Since $f(-t) = f(t)$, the probability curve is symmetrical about the line $t = 0$. As t increases, $f(t)$ decreases rapidly and tends to zero as $t \to \infty$, so that t-axis is an asymptote to the curve. We have shown that

$$\mu_2 = \frac{n}{n-2}, \; n > 2 \; ; \quad \beta_2 = \frac{3(n-2)}{(n-4)}, \; n > 4$$

Hence for $n > 2$, $\mu_2 > 1$ *i.e.*, the variance of t-distribution is greater than that of standard normal distribution and for $n > 4$, $\beta_2 > 3$ and thus t-distribution is more flat on the top than the normal curve. In fact, for small n, we have

$$P[\,|t| \geq t_0] \geq P[\,|Z| \geq t_0], \quad Z \sim N(0, 1)$$

i.e., the tails of the t-distribution have a greater probability (area) than the tails of standard normal distribution. Moreover we have also seen [§ 14·2·5] that for large n (*d.f.*), t-distribution tends to standard normal distribution.

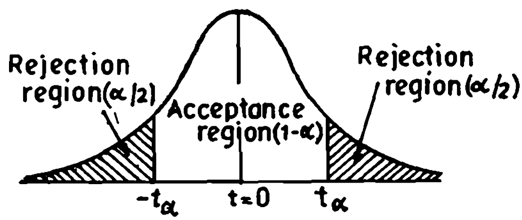


14·2·7. Critical Values of t. The critical (or significant) values of t at level of significance α and *d.f.* ν for *two-tailed test* are given by the equation

$$P[\,|t| > t_\nu(\alpha)] = \alpha \qquad \text{...(14·5)}$$

$$\Rightarrow \quad P[\,|t| \leq t_\nu(\alpha)] = 1 - \alpha \qquad \text{...(14·5a)}$$

CRITICAL VALUES OF t-DISTRIBUTION



The values $t_\nu(\alpha)$ have been tabulated in Fisher and Yates' Tables, for different values of α and ν and are given in the Appendix at the end of the book.

Since t-distribution is symmetric about $t = 0$, we get from (14·5)

$$P\,(t > t_\upsilon(\alpha)] + P\,[t < -t_\upsilon(\alpha)] = \alpha$$

$$\Rightarrow \quad 2P\,[t > t_\upsilon(\alpha)] = \alpha$$

$$\Rightarrow \quad P[t > t_\upsilon(\alpha)] = \alpha/2$$

$$\Rightarrow \quad P[t > t_\upsilon(2\alpha)] = \alpha \qquad \ldots(14\cdot5b)$$

$t_\upsilon(2\alpha)$ (from the Tables in the Appendix) gives the significant value of t for a single-tail test, [Right-tail or Left-tail-since the distribution is symmetrical], at level of significance α and ν *d.f.*

Hence the significant values of t at level of significance 'α' for a single tailed test can be obtained from those of two-tailed test by looking the values at level of significance '2α'.

For example,

t_8 (0·05) for single-tail test = t_8 (0·10) for two-tail test = 1·86

t_{15}(0·01) for single-tail test = t_{15} (0·02) for two-tail test = 2·60.

14·2·8. Applications of t-distribution. The t-distribution has a wide number of applications in Statistics, some of which are enumerated below.

(*i*) To test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean.

(*ii*) To test the significance of the difference between two sample means.

(*iii*) To test the significance of an observed sample correlation co-efficient and sample regression coefficient.

(*iv*) To test the significance of observed partial and multiple correlation coefficients.

In the following sections we will discuss these applications in detail, one by one.

14·2·9. t-Test for Single Mean. Suppose we want to test :

(*i*) if a random sample x_i ($i = 1, 2, \ldots, n$) of size n has been drawn from a normal population with a specified mean, say μ_0, or

(*ii*) if the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Under the null hypothesis H_0 :

(i) The sample has been drawn from the population with mean μ or (ii) there is no significant difference between the sample mean $\bar{x}$ and the population mean μ,

the statistic
$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \qquad \ldots(14\cdot6)$$

where
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \ldots[14\cdot6(a)]$$

follows Student's t-distribution with $(n-1)$ d.f.

We now compare the calculated value of t with the tabulated value at certain level of significance. If calculated $|t| >$ tabulated t, null hypothesis is rejected and if calculated $|t| <$ tabulated t, H_0 may be accepted at the level of significance adopted.

Remarks 1. *On computation of S^2 for numerical problems.* If $\bar{x}$ comes out in integers, the formula (14·6a) can be conveniently used for computing S^2. However, if $\bar{x}$ comes in fractions then the formula (14·6a) for computing S^2 is very cumbersome and is not recommended. In that case, step deviation method, given below, is quite useful.

If we take, $d_i = x_i - A$, where A is any arbitrary number then

$$S^2 = \frac{1}{n-1}\left[\sum(x_i - \bar{x})^2\right] = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] \qquad \ldots[14\cdot6(b)]$$

$$= \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right], \qquad \ldots[14\cdot6(c)]$$

since variance is independent of change of origin.

Also, in this case $\bar{x} = A + \dfrac{\sum d_i}{n}$. $\qquad \ldots[14\cdot6(d)]$

2. We know, the sample variance

$$s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$
$$ns^2 = (n-1)S^2$$
$$\Rightarrow \quad \frac{S^2}{n} = \frac{s^2}{n-1} \qquad \ldots[14\cdot6(e)]$$

Hence for numerical problems, the test statistic (14·6) on using [14·6(e)] becomes

$$t = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu_0}{\sqrt{s^2/(n-1)}} \sim t_{n-1} \qquad \ldots[14\cdot6(f)]$$

3. Assumptions for Student's t-test. The following assumptions are made in the Student's t-test :

(*i*) The parent population from which the sample is drawn is normal.

(*ii*) The sample observations are independent, *i.e.*, the sample is random.

(*iii*) The population standard deviation σ is unknown.

Example 14·2. *A machinist is making engine parts with axle diameters of 0·700 inch. A random sample of 10 parts shows a mean diameter of 0·742 inch with a standard deviation of 0·040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.*

Solution. Here we are given :

$$\mu = 0\cdot700 \text{ inches}, \bar{x} = 0\cdot742 \text{ inches}, s = 0\cdot040 \text{ inches and } n = 10$$

Null Hypothesis, $H_0 : \mu = 0\cdot700$, *i.e., the product is conforming to specifications.*

Alternative Hypothesis, $H_1 : \mu \neq 0\cdot700$

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{(n-1)}$$

Now

$$t = \frac{\sqrt{9}(0\cdot742 - 0\cdot700)}{0\cdot040} = 3\cdot15$$

How to proceed further. Here the test statistic 't' follows Student's t-distribution with $10 - 1 = 9$ d.f. We will now compare this calculated value with the tabulated value of t for 9 d.f. and at certain level of significance, say 5%. Let this tabulated value be denoted by t_0.

(*i*) If calculated 't' *viz.,* $3\cdot15 > t_0$, we say that the value of t is significant. This implies that $\bar{x}$ differs significantly from μ and H_0 is rejected at this level of significance and we conclude that the product is not meeting the specifications.

(*ii*) If calculated $t < t_0$, we say that the value of t is not significant, *i.e.,* there is no significant difference between $\bar{x}$ and μ. In other words, the deviation $(\bar{x} - \mu)$ is just due to fluctuations of sampling and null hypothesis H_0 may be retained at 5% level of significance, *i.e.,* we may take the product conforming to specifications.

Example 14·3. *The mean weekly sales of soap bars in departmental stores was 146·3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153·7 and showed a standard deviation of 17·2. Was the advertising campaign successful ?*

Solution. We are given : $n = 22, \bar{x} = 153\cdot7, s = 17\cdot2$.

Null Hypothesis. The advertising campaign is not successful, *i.e.,*

$$H_0 : \mu = 146\cdot3$$

Alternative Hypothesis. $H_1 : \mu > 146\cdot3$ (Right-tail).

Test Statistic. Under the null hypothesis, the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{22-1} = t_{21}$$

Now

$$t = \frac{153\cdot7 - 146\cdot3}{\sqrt{(17\cdot2)^2/21}} = \frac{7\cdot4 \times \sqrt{21}}{17\cdot2} = 9\cdot03$$

Conclusion. Tabulated value of t for 21 *d.f.* at 5% level of significance for *single-tailed test* is 1·72. Since calculated value is much greater than the tabulated value, it is highly significant. Hence we reject the null hypothesis and conclude that the advertising campaign was definitely successful in promoting sales.

Example 14·4. *A random sample of 10 boys had the following I.Q.'s : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100 ? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.*

[***Madras Univ. B.E., April 1990***]

Solution. *Null hypothesis, H_0 : The data are consistent with the assumption of a mean I.Q. of 100 in the population, i.e., $\mu = 100$.*

Alternative hypothesis, $H_1 : \mu \neq 100$.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{(\bar{x} - \mu)}{\sqrt{S^2/n}} \sim t_{(n-1)},$$

where $\bar{x}$ and S^2 are to be computed from the sample values of I.Q.'s.

CALCULATIONS FOR SAMPLE MEAN AND S.D.

X	$(X - \bar{x})$	$(X - \bar{x})^2$
70	−27·2	739·84
120	22·8	519·84
110	12·8	163·84
101	3·8	14·44
88	−9·2	84·64
83	−14·2	201·64
95	−2·2	4·84
98	0·8	0·64
107	9·8	96·04
100	2·8	7·84
Total 972		1833·60

Hence $n = 10, \bar{x} = \frac{972}{10} = 97{\cdot}2$ and $S^2 = \frac{1833{\cdot}60}{9} = 203{\cdot}73$

$$\therefore \qquad |t| = \frac{|97{\cdot}2 - 100|}{\sqrt{203{\cdot}73/10}} = \frac{2{\cdot}8}{\sqrt{20{\cdot}37}} = \frac{2{\cdot}8}{4{\cdot}514} = 0{\cdot}62$$

Tabulated $t_{0{\cdot}05}$ for $(10 - 1)$ *i.e.,* 9 d.f. for two-tailed test is 2·262.

Conclusion. Since calculated t is less than tabulated $t_{0{\cdot}05}$ for 9 d.f., H_0 may be accepted at 5% level of significance and we may conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population.

The 95% confidence limits within which the mean I.Q. values of samples of 10 boys will lie are given by

$$\bar{x} \pm t_{0{\cdot}05}\, S/\sqrt{n} = 97{\cdot}2 \pm 2{\cdot}262 \times 4{\cdot}514$$

$$= 97{\cdot}2 \pm 10{\cdot}21 = 107{\cdot}41 \text{ and } 86{\cdot}99$$

Hence the required 95% confidence interval is [86·99, 107·41].

Remark. Aliter for computing $\bar{x}$ and S^2. Here we see that $\bar{x}$ comes in fractions and as such the computation of $(x-\bar{x})^2$ is quite laborious and time consuming. In this case we use the method of step deviations to compute $\bar{x}$ and S^2, as given below.

X	$d = X - 90$	d^2
70	−20	400
120	30	900
110	20	400
101	11	121
88	−2	4
83	−7	49
95	5	25
98	8	64
107	17	289
100	10	100
Total	$\Sigma d = 72$	$\Sigma d^2 = 2352$

Here $\quad d = X - A$, where $A = 90$

$$\therefore \quad \bar{x} = A + \frac{1}{n}\Sigma d = 90 + \frac{72}{10} = 97{\cdot}2$$

and
$$S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{9}\left[2352 - \frac{(72)^2}{10}\right] = 203{\cdot}73$$

Example 14·5. *The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches ? Test at 5% significance level, assuming that for 9 degrees of freedom $P(t > 1{\cdot}83) = 0{\cdot}05$.*

Solution. *Null Hypothesis,* $H_0 : \mu = 64$ inches.

Alternative Hypothesis, $H_1 : \mu > 64$ inches.

CALCULATIONS FOR SAMPLE MEAN AND S.D.

x	70	67	62	68	61	68	70	64	64	66	Total 660
$x-\bar{x}$	4	1	−4	2	−5	2	4	−2	−2	0	0
$(x-\bar{x})^2$	16	1	16	4	25	4	16	4	4	0	90

$$\bar{x} = \frac{\Sigma x}{n} = \frac{660}{10} = 66$$

$$S^2 = \frac{1}{n-1}\Sigma(x-\bar{x})^2 = \frac{90}{9} = 10$$

Test Statistic. Under H_0, the test statistic is

$$t = \frac{\bar{x}-\mu}{\sqrt{S^2/n}} = \frac{66-64}{\sqrt{10/10}} = 2,$$

which follows Student's t-distribution with $10 - 1 = 9$ *d.f.*

Tabulated value of t for 9 *d.f.* at 5% level of significance for single (right) tail-test is 1·833. (This is the value $t_{0.10}$ for 9 *d.f.* in the two-tailed Table given in the Appendix.)

Conclusion. Since calculated value of t is greater than the tabulated value, it is significant. Hence H_0 is rejected at 5% level of significance and we conclude that the average height is greater than 60 inches.

Example 14·6. *A random sample of 16 values from a normal population showed a mean of 41·5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43·5 inches for the population is not reasonable. Obtain 95 per cent and 99 per cent fiducial limits for the same.*

You may use the following information from statistical tables :

$$v = 15, \begin{cases} P = 0.05, \ t = 2.131 \\ P = 0.01, \ t = 2.947 \end{cases}$$

Solution. We are given $n = 16$, $\bar{x} = 41·5$ inches and

$\Sigma(x-\bar{x})^2 = 135$ sq. inches.

$$\therefore \qquad S^2 = \frac{1}{n-1}\Sigma(x-\bar{x})^2 = \frac{135}{15} = 9 \quad \Rightarrow \quad S = 3$$

Null Hypothesis, $H_0: \mu = 43·5$ inches, *i.e.*, the data are consistent with the assumption that the mean height in the population is 43·5 inches.

Alternative Hypothesis, $H_1: \mu \neq 43·5$ inches.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

Now
$$|t| = \frac{|41·5 - 43·5|}{3/4} = \frac{8}{3} = 2·667$$

Here number of degrees of freedom is $(16 - 1) = 15$.

We are given :

$t_{0.05}$ for 15 d.f. = 2·131 and $t_{0.01}$ for 15 d.f. = 2·947

Conclusion. Since calculated $|t|$ is greater than 2·131, null hypothesis is rejected at 5% level of significance and we conclude that the assumption of mean of 43·5 inches for the population is not reasonable.

Remark. Since calculated $|t|$ is less than 2·947, null hypothesis ($\mu = 43·5$) may be accepted at 1% level of significance.

95% fiducial limits for μ : (d.f. = 15)

$$\bar{x} \pm t_{0.05} \times \frac{S}{\sqrt{n}} = 41{\cdot}5 \pm 2{\cdot}131 \times \frac{3}{4} = 41{\cdot}5 \pm 1{\cdot}598$$

$\therefore$ $$39{\cdot}902 < \mu < 43{\cdot}098$$

99% fiducial limits for μ : ($d.f.$ = 15)

$$\bar{x} \pm t_{0.01} \times \frac{S}{\sqrt{n}} = 41{\cdot}5 \pm 2{\cdot}947 \times \frac{3}{4} = 43{\cdot}71 \text{ and } 39{\cdot}29$$

$\therefore$ $$39{\cdot}29 < \mu < 43{\cdot}71$$

EXERCISE 14(b)

1. (*a*) Write a short note on Student's t-distribution and point out its uses.

(*b*) Show how the t-distribution has been found useful in testing whether the mean of small sample is significantly different from a hypothetical value.

(*c*) It is desired to test the hypothesis that the mean of a normal population is $\mu = \mu_0$ against the alternative that $\mu \neq \mu_0$. Explaining the assumptions involved, develop the statistic suitable for testing this hypothesis if the size of the sample is small. What modification do you suggest when the sample size is large ?

2. What is a test of significance ?

To test the hypothesis that the mean of a normal distribution is zero, two independent observations x_1 and x_2 are taken from the distribution. Show that the hypothesis is rejected at 10% level of significance, using t test with equal tail ends, if

$$| x_1 + x_2 | > | x_1 - x_2 | \tan 81^0$$

3. It is required to test that the mean of a normal population is zero. A random sample drawn from the population gives the values $x_1, x_2, \ldots, x_n$. Show that the t-test for acceptance of the hypothesis reduces to

$$\left(\sum_{i=1}^{n} x_i\right) \leq \frac{n \cdot t_\alpha^2}{t_\alpha^2 + (n-1)} \left(\sum_{i=1}^{n} x_i^2\right)$$

where t_α is the value of Student's t at the desired level of significance α for $(n-1)$ $d.f.$

4. (*a*) Find the Student's t for following variate values in a sample of eight : –4, –2, –2, 0, 2, 2, 3, 3, taking the mean of the universe to be zero. How would you proceed further ?

(*b*) Ten individuals are chosen at random from a normal population and their heights are found to be 63, 63, 66, 67, 68, 69 70, 70, 71, 71 inches. Test if the sample belongs to the population whose mean heights is 66″

[Given $t_{0.05} = 2{\cdot}62$ for 9 $d.f.$]

(*c*) A random sample of 9 experimental animals under a certain diet gave the following increase in weight : $\Sigma x_i = 45$ lbs, $\Sigma x_i^2 = 279$ lbs., where x_i denotes the increase in weight of the ith animal. Assuming that the increase in weight is normally distributed as $N(\mu, \sigma^2)$ variate, test $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$ at 5% level. Given $P(|t| > 2{\cdot}306) = 0{\cdot}05$ for 8 degrees of freedom.

[*Calcutta Univ. B.Sc. (Maths.Hons.), 1991*]

5. A manufacturer of gunpowder has developed a new powder which is designed to produce a muzzle velocity equal to 3000 ft/sec. Seven shells are loaded with the charge and the muzzle velocities measured.

The resulting velocities are as follows : 3,005; 2,935; 2,965; 2,995; 3,905, 2,935; and 2,905. Do these data present sufficient evidence to indicate that the average velocity differs from 3,000 ft./sec.

6. The average length of time for students to register for summer classes at a certain college has been 50 minutes with a standard deviation of 10 minutes. A new registration procedure using modern computing machines is being tried. If a random sample of 12 students had an average registration time of 42 minutes with s.d. of 11·9 minutes under the new system, test the hypothesis that the population mean has not changed, using ·05 as level of significance.

7. The nine items of a sample had the following values : 45, 47, 50, 52, 48, 47, 49, 53 and 51.

Does the mean of the nine items differ significantly from the assumed population mean of 47·5 ? Given that

$$\nu = 8, \begin{cases} P = 0 \cdot 945 \text{ for } t = 1 \cdot 8 \\ P = 0 \cdot 953 \text{ for } t = 1 \cdot 9 \end{cases}$$

8. A time study engineer developed a new sequence of operation elements that he hopes will reduce the mean cycle time of a certain production process. The results of a time study of 20 cycles are given below :

cycle time in minutes

12·25	11·97	12·15	12·08	12·31	12·28	11·94	11·89	12·16	12·04
12·09	12·15	12·14	12·47	11·98	12·04	12·11	12·25	12·15	12·34

If the present mean cycle time is 12·5 minutes, should he adopt the new sequence ?

9. (*a*) The average breaking strength of steel rods is specified to be 18·5 thousand pounds. To test this a sample of 14 rods was tested. The mean and standard deviations obtained were 17·85 and 1·955 thousand pounds respectively. Is the result of the experiment significant ? Also obtain the 95 per cent fiducial limits from the sample for the average breaking strength of steel rods.

(*b*) A sample of 9 shafts is inspected from a production line. The following measurements are the diameters (in mm.) of shafts : 45·010, 45·020, 45·021, 45·015, 45·019, 45·018, 45·020, 45·023 and 45·005. If the production line meets the specifications laid by the I.S.I., with S.D. 0·006 mm, estimate the 95% confidence interval within which the true diameter of the shaft lies.

[*Madras Univ. B.E., 1989*]

10. a random sample of 8 envelopes is taken from letter box of a post office and their weights in grams are found to be 12·1, 11·9, 12·4, 12·3, 11·9, 12·1, 12·4, 12·1.

(*a*) Find 99% confidence limits for the mean weight of the envelopes received at that post office.

(*b*) Using the result of part (*a*), does this sample indicate at 1% level that the average weight of envelopes received at that post office is 12·35 gms.

11. A random sample of nine from men of a large city gave a mean height 68 inches and the unbiased estimate of the population variance found from the

sample was 4·5 inches. Proceed as far as you can to test for a mean height of 68·5 inches for the men of the city. Also state how you would proceed further.

14·2·10. t-Test for Difference of Means. Suppose we want to test if two independent samples x_i ($i = 1, 2, \ldots, n_1$) and y_j, ($j = 1, 2, \ldots, n_2$) of sizes n_1 and n_2 have been drawn from two normal populations with means μ_X and μ_Y respectively.

Under the null hypothesis (H_0) *that the samples have been drawn from the normal populations with means* μ_X *and* μ_Y *and under the assumption that the population variance are equal, i.e.,* $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (say), the statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \qquad \ldots(14\cdot7)$$

where $\quad \bar{x} = \dfrac{1}{n_1}\sum\limits_{i=1}^{n_1} x_i, \qquad \bar{y} = \dfrac{1}{n_2}\sum\limits_{j=1}^{n_2} y_j$

and $\quad S^2 = \dfrac{1}{n_1 + n_2 - 2}\left[\sum\limits_i (x_i - \bar{x})^2 + \sum\limits_j (y_j - \bar{y})^2\right] \qquad \ldots[14\cdot7(a)]$

is an unbiased estimate of the common population variance σ^2, follows Student's t-distribution with $(n_1 + n_2 - 2)$ d.f.

Proof. *Distribution of t defined in (14·7).*

$$\xi = \frac{(\bar{x} - \bar{y}) - E(\bar{x} - \bar{y})}{\sqrt{V(\bar{x} - \bar{y})}} \sim N(0, 1)$$

But $\quad E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_X - \mu_Y$

and $\quad V(\bar{x} - \bar{y}) = V(\bar{x}) + V(\bar{y})$

[The covariance term vanishes since samples are independent.]

$$= \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad \text{(By assumption)}$$

$$\therefore \quad \xi = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0, 1) \qquad \ldots(*)$$

Let $\quad \chi^2 = \left[\sum\limits_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum\limits_{j=1}^{n_2} (y_j - \bar{y})^2\right] \Big/ \sigma^2$

$$= \left[\sum_i (x_i - \bar{x})^2/\sigma^2\right] + \left[\sum_j (y_j - \bar{y})^2/\sigma^2\right] = \frac{n_1 s_X^2}{\sigma^2} + \frac{n_2 s_Y^2}{\sigma^2} \qquad \ldots(**)$$

Since $n_1 s_X^2/\sigma^2$ and $n_2 s_Y^2/\sigma^2$ are independent χ^2-variates with $(n_1 - 1)$ and $(n_2 - 1)$ d.f. respectively, by the additive property of chi-square distribution, χ^2 defined in (**) is a χ^2-variate with $(n_1 - 1) + (n_2 - 1)$, *i.e.*, $n_1 + n_2 - 2$ *d.f.*

Further, since sample mean and sample variance are independently distributed, ξ and χ^2 are independent random variables.

Hence Fisher's t statistic is given by

$$t = \frac{\xi}{\sqrt{\dfrac{\chi^2}{n_1 + n_2 - 2}}}$$

$$= \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

$$\times \frac{1}{\left[\dfrac{1}{n_1 + n_2 - 2}\left\{\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right\}\Big/\sigma^2\right]^{1/2}}$$

$$= \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}},$$

where $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right]$$

and it follows Student's t-distribution with $(n_1 + n_2 - 2)$ d.f. (*c.f.* Remark § 14·2·3, page 14·4).

Remarks 1. S^2, defined in 14·7(a) is an unbiased estimate of the common population variance σ^2, since

$$E(S^2) = \frac{1}{n_1 + n_2 - 2} E\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right]$$

$$= \frac{1}{n_1 + n_2 - 2} E\left[(n_1 - 1) S_X^2 + (n_2 - 1) S_Y^2\right]$$

$$= \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1) E(S_X^2) + (n_2 - 1) E(S_Y^2)\right]$$

$$= \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2\right] = \sigma^2$$

2. An important deduction which is of much practical utility is discuss[ed] below :

Suppose we want to test if : (a) two independent samples x_i ($i = 1, 2, \ldots, n_1$), and y_j ($j = 1, 2, \ldots, n_2$), have been drawn from the populations with sa[me] means or (b) the two sample means $\bar{x}$ and $\bar{y}$ differ significantly or not.

Under the null hypothesis H_0 that (a) *samples have been drawn from t[he] populations with the same means, i.e., $\mu_X = \mu_Y$ or (b) the sample means $\bar{x}$ a[nd] $\bar{y}$ do not differ significantly*, [From (14·7)] the statistic :

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad [\because \mu_X = \mu_Y, \text{ under } H_0] \qquad ...(14\cdot8)$$

where symbols are defined in (14·7*a*),follows Student's *t*-distribution with $(n_1 + n_2 - 2)$ d.f.

3. *On the assumption of t-test for difference of means.* Here we make the following three fundamental assumptions :

(*i*) Parent populations, from which the samples have been drawn are normally distributed.

(*ii*) The population variances are equal and unknown, *i.e.*, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, (say), where σ^2 is unknown.

(*iii*) The two samples are random and independent of each other.

Thus before applying *t*-test for testing the equality of means it is theoretically desirable to test the equality of population variances by applying *F*-test. If the variances do not come out to be equal then *t*-test becomes invalid and in that case Behren's '*d*'-test based on fiducial intervals is used. For practical problems, however, the assumptions (*i*) and (*ii*) are taken for granted.

4. Paired t-test For Difference of Means. Let us now consider the case when (*i*) the sample sizes are equal, *i.e.*, $n_1 = n_2 = n$ (say), and (*ii*) the two samples are not independent but the sample observations are paired together, *i.e.*, the pair of observations (x_i, y_i), $(i = 1, 2, ..., n)$ corresponds to the same (*i*th) sample unit. The problem is to test if the sample means differ significantly or not.

For example,suppose we want to test the efficacy of a particular drug, say, for inducing sleep. Let x_i and y_i $(i = 1, 2, ..., n)$ be the readings, in hours of sleep, on the *i*th individual, before and after the drug is given respectively. Here instead of applying the difference of the means test discussed in § 14·2·10, we apply the paired *t*-test given below.

Here we consider the increments, $d_i = x_i - y_i$, $(i = 1, 2, ..., n)$.

Under the null hypothesis, H_0 that *increments are due to fluctuations of sampling, i.e., the drug is not responsible for these increments,* the statistic.

$$t = \frac{\bar{d}}{S/\sqrt{n}} \qquad ...(14\cdot9)$$

where $$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 \qquad ...[14\cdot9(a)]$$

follows Student's *t*-distribution with $(n - 1)$ d.f.

Example 14·7. *Below are given the gain in weights (in lbs.) of pigs fed on two diets A an B.*

Gain in weight

Diet A : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet B : 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test, if the two diets differ significantly as regards their effect on increase in weight.

Solution. *Null hypothesis,* $H_0 : \mu_X = \mu_Y$, *i.e., there is no significant difference between the mean increase in weight due to diets A and B.*

Alternative hypothesis, $H_1: \mu_X \neq \mu_Y$ (two-tailed).

Diet A

	X	$X-\bar{X}$	$(X-\bar{X})^2$
	25	−3	9
	32	4	16
	30	2	4
	34	6	36
	24	−4	16
	14	−14	196
	32	4	16
	24	−4	16
	30	2	4
	31	3	9
	35	7	49
	25	−3	9
Total	336	0	380

Diet B

	Y	$Y-\bar{Y}$	$(Y-\bar{Y})^2$
	44	14	196
	34	4	16
	22	−8	64
	10	−20	400
	47	17	289
	31	1	1
	40	10	100
	30	0	0
	32	2	4
	35	5	25
	18	−12	144
	21	−9	81
	35	5	25
	29	−1	1
	22	−8	64
Total	450	0	1410

Under null hypothesis (H_0) :

$$t = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

Here $\left.\begin{array}{r} n_1 = 12, \\ \Sigma x = 336 \\ \Sigma(x-\bar{x})^2 = 380 \end{array}\right\}$ and $\left.\begin{array}{r} n_2 = 15 \\ \Sigma y = 450 \\ \Sigma(y-\bar{y})^2 = 1410 \end{array}\right\}$

$\therefore$ $$\bar{x} = \frac{336}{12} = 28, \quad \bar{y} = \frac{450}{15} = 30$$

and $$S^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2\right] = 71{\cdot}6$$

$\therefore$ $$t = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} = \frac{28-30}{\sqrt{71{\cdot}6\left(\dfrac{1}{12}+\dfrac{1}{15}\right)}}$$

$$= \frac{-2}{\sqrt{10 \cdot 74}} = -0 \cdot 609$$

Tabulate $t_{0.05}$ for $(12 + 15 - 2) = 25$ d.f. is 2·06.

Conclusion. Since calculated $|t|$ is less than tabulated t, H_0 may be accepted at 5% level of significance and we may conclude that the two diets do not differ significantly as regards their effect on increase in weight.

Remark. Here $\bar{x}$ and $\bar{y}$ come out to be integral values and hence the direct method of computing $\Sigma(x - \bar{x})^2$ and $\Sigma(y - \bar{y})^2$ is used. In case $\bar{x}$ and (or) $\bar{y}$ comes out to be fractional, then the step deviation method is recommended for computation of $\Sigma(x - \bar{x})^2$ and $\Sigma(y - \bar{y})^2$.

Example 14·8. *Samples of two types of electric light bulbs were tested for length of life and following data were obtained :*

	Type I	*Type II*
Sample No.	$n_1 = 8$	$n_2 = 7$
Sample Means	$\bar{x}_1 = 1,234$ *hrs.*	$\bar{x}_2 = 1,036$ *hrs.*
Sample S.D.'s	$s_1 = 36$ *hrs.*	$s_2 = 40$ *hrs.*

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life ?

Solution. *Null Hypothesis,* $H_0 : \mu_X = \mu_Y$, *i.e.,* the two types I and II of electric bulbs are indentical.

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$, *i.e.,* type I is superior to type II.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2} = t_{13},$$

where
$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2\right]$$

$$= \frac{1}{n_1 + n_2 - 2}\left[n_1 s_1^2 + n_2 s_2^2\right] = \frac{1}{13}\left[8 \times (36)^2 + 7 \times (40)^2\right] = 1659 \cdot 08$$

$$\therefore \quad t = \frac{1234 - 1036}{\sqrt{1659 \cdot 08\left(\frac{1}{8} + \frac{1}{7}\right)}} = \frac{198}{\sqrt{1659 \cdot 08 \times 0 \cdot 2679}} = 9 \cdot 39$$

Tabulated value of t for 13 *d.f.* at 5% level of significance for right (single) tailed test is 1·77. [This is the value of $t_{0.10}$ for 13 *d.f.* from two-tail tables given in Appendix].

Conclusion. Since calculated 't' is much greater than tabulated 't', it is highly significant and H_0 is rejected. Hence the two types of electric bulbs differ significantly. Further since $\bar{x}_1$ is much greater than $\bar{x}_2$, we conclude that type I is definitely superior to type II.

Example 14·9. *The heights of six randomly chosen sailors are in inches : 63, 65, 68, 69, 71, and 72. Those of 10 randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss, the light that these data throw on the suggestion that sailors are on the average taller than soldiers.*

Solution. If the heights of sailors and soldiers be represented by the variables X and Y respectively then the *Null Hypothesis* is, $H_0 : \mu_X = \mu_Y$, *i.e., the sailors are not on the average taller than the soldiers.*

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$ (Right-tailed).

Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{14}$$

Sailors X	$d = X - A$ $= X - 68$	d^2	*Soldiers* Y	$D = Y - B$ $= Y - 66$	D^2
63	−5	25	61	−5	25
65	−3	9	62	−4	16
68	0	0	65	−1	1
69	1	1	66	0	0
71	3	9	69	3	9
72	4	16	69	3	9
			70	4	16
Total	0	60	71	5	25
			72	6	36
			73	7	49
			Total	18	186

$$\therefore \quad \bar{x} = A + \frac{\sum d}{n_1} = 68 + 0 = 68$$

and $$\sum(x - \bar{x})^2 = \sum d^2 - \frac{(\sum d)^2}{n_1} = 60 - 0 = 60$$

$$\bar{y} = B + \frac{\sum D}{n_2} = 66 + \frac{18}{10} = 67{\cdot}8$$

and $$\sum(y - \bar{y})^2 = \sum D^2 - \frac{(\sum D)^2}{n_2} = 186 - \frac{324}{10} = 153{\cdot}6$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2\right] = \frac{1}{14}(60 + 153{\cdot}6) = 15{\cdot}2571$$

$$\therefore \quad t = \frac{68 - 67\cdot8}{\sqrt{15\cdot2571}\left(\frac{1}{6} + \frac{1}{10}\right)^{1/2}} = \frac{0\cdot2}{\sqrt{15\cdot2571 \times 0\cdot2667}} = 0\cdot099$$

Tabulated $t_{0\cdot05}$ for 14 $d.f.$ for single-tail test is 1·76.

Conclusion. Since calculated t is much less than 1·76, it is not at all significant at 5% levels of significance. Hence null hypothesis may be retained at 5% level of significance and we conclude that the data are inconsistent with the suggestion that the sailors are on the average taller than soldiers.

Example 14·10. *A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure :*

5, 2, 8, –1, 3, 0, –2, 1, 5, 0, 4 and 6

Can it be concluded that the stimulus will, in general, be accompanied by an increase in blood pressure ? [*Delhi Univ. B.Sc. 1989*]

Solution. Here we are given the increments in blood pressure *i.e.,* $d_i \ (= x_i - y_i)$.

Null Hypothesis, $H_0 : \mu_X = \mu_Y$, *i.e.,* there is no significant difference in the blood pressure readings of the patients before and after the drug. In other words, the given increments are just by chance (fluctuations of sampling) and not due to the stimulus.

Alternative Hypothesis, $H_1 : \mu_X < \mu_Y$, *i.e.,* the stimulus results is an increase in blood pressure.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

d	5	2	8	–1	3	0	–2	1	5	0	4	6	31
d^2	25	4	64	1	9	0	4	1	25	0	16	36	185

$$S^2 = \frac{1}{n-1}\Sigma(d - \bar{d})^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right]$$

$$= \frac{1}{11}\left[185 - \frac{(31)^2}{12}\right] = \frac{1}{11}(185 - 80\cdot08) = 9\cdot5382$$

and $$\bar{d} = \frac{\Sigma d}{n} = \frac{31}{12} = 2\cdot58$$

$$\therefore \quad t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2\cdot58 \times \sqrt{12}}{\sqrt{9\cdot5382}} = \frac{2\cdot58 \times 3\cdot464}{3\cdot09} = 2\cdot89$$

Tabulated $t_{0\cdot05}$ for 11 d.f. for right-tail test is 1·80. [This is the value of $t_{0\cdot10}$ for 11 $d.f$ in the Table for two-tailed test given in the Appendix].

Conclusion. Since calculated $t > t_{0.05}$, H_0 is rejected at 5% level of significance. Hence we conclude that the stimulus will, in general, be accompanied by an increase in blood pressure.

Example 14·11. *In a certain experiment to compare two types of pig foods A and B, the following results of increase in weights were observed in pigs :*

Pig number		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
Increase in weight in lb	*Food A*	49	53	51	52	47	50	52	53	407
	Food B	52	55	52	53	50	54	54	53	423

(i) Assuming that the two samples of pigs are independent, can we conclude that food B is better than food A ?

(ii) Also examine the case when the same set of eight pigs were used in both the foods.

Solution. *Null Hypothesis,* H_0. If the increase in weights due to foods A and B are denoted by X and Y respectively then $H_0 : \mu_X = \mu_Y$, *i.e.*, there is no significant difference in increase in weights due to diets A and B.

Alternative Hypothesis, $H_1 : \mu_X < \mu_Y$ (Left-tailed).

(*i*) If the two samples of pigs be assumed to be independent, then we will apply t-test for difference of means to test H_0.

Test Statistic. Under $H_0 : \mu_X = \mu_Y$, the test criterion is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

Food A			*Food B*		
X	$d = X - 50$	d^2	Y	$D = Y - 52$	D^2
49	−1	1	52	0	0
53	3	9	55	3	9
51	1	1	52	0	0
52	2	4	53	1	1
47	−3	9	50	−2	4
50	0	0	54	2	4
52	2	4	54	2	4
53	3	9	53	1	1
	7	37		7	23

$$\therefore \bar{x} = 50 + \frac{7}{8} = 50{\cdot}875 \qquad \bar{y} = 52 + \frac{7}{8} = 52{\cdot}875$$

and $\left.\begin{aligned}\Sigma(x-\bar{x})^2 &= \Sigma d^2 - \frac{(\Sigma d)^2}{n_1}\\ &= 37 - \frac{49}{8}\\ &= 30{\cdot}875\end{aligned}\right\}$ $\left.\begin{aligned}\Sigma(y-\bar{y})^2 &= \Sigma D^2 - \frac{(\Sigma D)^2}{n_2}\\ &= 23 - \frac{49}{8}\\ &= 16{\cdot}875\end{aligned}\right\}$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2\right]$$

$$= \frac{1}{14}(30{\cdot}875 + 16{\cdot}875) = 3{\cdot}41$$

$$\therefore \qquad t = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{50{\cdot}875 - 52{\cdot}875}{\sqrt{3{\cdot}41\left(\frac{1}{8}+\frac{1}{8}\right)}} = -2{\cdot}17$$

Tabulated $t_{0{\cdot}05}$ for $(8 + 8 - 2) = 14$ d.f. for one-tail test is 1·76.

Conclusion. The critical region for the left-tail test is $t < -1{\cdot}76$. Since calculated t is less than $-1{\cdot}76$, H_0 is rejected at 5% level of significance. Hence we conclude that the foods A and B differ significantly as regards their effect on increase in weight. Further, since $\bar{y} > \bar{x}$, food B is superior to food A.

(*ii*) If the same set of pigs is used in both the cases, then the readings X and Y are not independent but they are paired together and we apply the paired t-test for testing H_0.

Under $H_0 : \mu_X = \mu_Y$, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

X	49	53	51	52	47	50	52	53	Total
Y	52	55	52	53	50	54	54	53	
$d = X - Y$	−3	−2	−1	−1	−3	−4	−2	0	−16
d^2	9	4	1	1	9	16	4	0	44

$$\therefore \qquad \bar{d} = \frac{\Sigma d}{n} = \frac{-16}{8} = -2$$

and $$S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{7}\left[44 - \frac{256}{8}\right] = 1{\cdot}714$$

$$\therefore \qquad |t| = \frac{|\bar{d}|}{\sqrt{S^2/n}} = \frac{2}{\sqrt{1{\cdot}7143/8}} = \frac{2}{0{\cdot}4629} = 4{\cdot}32$$

Tabulated $t_{0{\cdot}05}$ for $(8 - 1) = 7$ d.f. for one-tail test is 1·90.

Conclusion. Here also the observed value of 't' is significant at 5% level of significance and we conclude that food B is superior to food A.

EXERCISE 14 (c)

1. Explain, stating clearly the assumptions involved, the t-test for testing the significance of the difference between the two sample means.

2. Two independent samples of 8 and 7 items respectively had the following values

Sample I...	9	11	13	11	15	9	12	14
Sample II...	10	12	10	14	9	8	10	

Is the difference between the means of samples significant ?

3. (*a*) Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results :

Horse A :	28	30	32	33	33	29	34
Horse B :	29	30	30	24	27	29	

Test whether the two horses have the same running capacity. [5 per cent values of t for 11 and 12 degrees of freedom respectively are 2·20 and 2·18].

Ans. Calculated $t = 2{\cdot}5$ (approx.)

(*b*) The gain in weight of two random samples of rats fed on two different diets A and B are given below. Examine whether the difference in mean increases in weight is significant.

Diet A :	13	14	10	11	12	16	10	8	
Diet B :	7	10	12	8	10	11	9	10	11

4. (*a*) Show how you would use Student's t-test to decide whether the two sets of observations

[17, 27, 18, 25,27, 29, 27, 23, 17] and [16,16, 20, 16, 20, 17, 15, 21]

indicate samples drawn from the same universe.

(*b*) A reading test is given to an elementary school class that consists of 12 Anglo-American children and 10 Mexican-American children. The results of the test are :

Anglo-American	*Mexican-American*
$\bar{x}_1 = 74$	$\bar{x}_2 = 70$
$s_1 = 8$	$s_2 = 10$

Is the difference between the means of the two groups significant at the 0·05 level ? Given $t_{20} = 2{\cdot}086$, $t_{22} = 2{\cdot}074$ at 5% level.

[*Delhi Univ. M.C.A., 1986*]

5. (*a*) For a random sample of 10 pigs, fed on a diet A, the increases in weight in a certain period were 10, 6, 16,17, 13, 12, 8, 14, 15, 9 lbs.

For another random sample of 12 pigs fed on diet B, the increases in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 lbs.

Find if the two samples are significantly different regarding the effect of diet, given that for d.f. ν = 20, 21, 22, the five per cent values of t are respectively 2·09, 2·07, 2·06.

Ans. $t = 1{\cdot}51$; Sample means do not differ significantly.

(*b*) Two independent samples of rats chosen among both the series had the following increase in weights when fed on a diet. Can you say that the mean increase in weight differs significantly with sex ?

Male :	96,	88,	97,	89,	92,	95	and	90		
Female :	112,	80,	98,	100,	84,	82,	89,	95,	100 and 96.	

6. (*a*) Ten soldiers visit a riffle range for two consecutive weeks. For the first week their scores are

67, 24, 57, 55, 63, 54, 56, 68, 33, 43

and during the second week they score in the same order—

70, 38, 58, 58, 56, 67, 68, 72, 42, 38

Examine if there is any significant difference in their performance.

(*b*) Two independent groups of 10 children were tested to find how many digits they could repeat from memory after hearing them. The results are as follows :

Group A :	8	6	5	7	6	8	7	4	5	6
Group B :	10	6	7	8	6	9	7	6	7	7

Is the difference between the mean scores of the two groups significant ?

(*c*) Measurements of the fat content of two kinds of ice cream, Brand *A* and Brand *B*, yielded the following sample data :

Brand A :	13·5	14·0	13·6	12·9	13·0
Brand B :	12·9	13·0	12·4	13·5	12·7

Test the null hypothesis $\mu_1 = \mu_2$, (where μ_1 and μ_2 are the respective true average fat contents of the two kinds of ice cream), against the alternative hypothesis $\mu_1 \neq \mu_2$ at the level of significance $\alpha = 0{\cdot}05$.

[*Madras Univ. B.E.*, *1990*]

7. (*a*) A random sample of 16 values from a normal population has a mean of 41·5 inches and sum of squares of deviations from the mean is equal to 135 inches. Another sample of 20 values form an unknown population has a mean of 43·0 inches and sum of squares of deviations from their mean is equal to 171 inches. Show that the two samples may be regarded as coming from the same normal population.

(*b*) A company is interested in knowing if there is a difference in the average salary received by foremen in two divisions. Accordingly samples of 12 foremen in the first division and 10 foremen in the second division are selected at random. Based upon experience, foremen's salaries are known to be approximately normally distributed, and the standard deviations are about the same.

	First Division	*Second division*
Sample size	12	10
Average monthly salary of foremen (Rs.)	1,050	980
Standard deviation of salaries (Rs.)	68	74

The table value of t for 20 d.f. at 5% level of significance is 2·086.

Ans. $t = 2{\cdot}2$. Reject $H_0 : \mu_X = \mu_Y$

(*c*) The average number of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 1% level of significance ?

Ans. $t = -7·65$. **Hint.** Here $n_1 = n_2 = 25$.

8. Eleven school boys were given a test in Statistics. They were given a month's tuition and a second test was held at the end of it. Do the marks give evidence that the students have benefited by the extra coaching ?

Boys	1	2	3	4	5	6	7	8	9	10	11
Marks in 1st test	23	20	19	21	18	20	18	17	23	16	19
Marks in 2nd test	24	19	22	18	20	22	20	20	23	20	18

Ans. $H_0: \mu_1 = \mu_2; H_1 : \mu_1 < \mu_2$. Paired $| t | = 1·483$. Not significant. Hence, students have not benefited from extra coaching.

9. (*a*) The following table gives the additional hours of sleep gained by 10 patients in an experiment to test the effect of a drug. Do these data give evidence that the drug produces additional hours of sleep ?

Patients	1	2	3	4	5	6	7	8	9	10
Hours gained :	0·7	0·1	0·2	1·2	0·31	0·4	3·7	0·8	3·8	2·0

(*b*) A drug was administered to 10 patients, and the increments in their blood pressure were recorded to be 6, 3, –2, 4, –3, 4, 6, 0, 3, 2. Is it reasonable to believe that the drug has no effect on change of blood pressure ? Use 5% significance level, and assume that for 9 degrees of freedom, $P(t > 2·26) = 0·025$. **[Calcutta Univ. B.Sc.(Maths. Hons.), 1986]**

(*c*) The scores of 10 candidates prior and after training are given below :

Prior :	84	48	36	37	54	69	83	96	90	65
After :	90	58	56	49	62	81	84	86	84	75

Is the training effective ? **[Calicut Univ. B.Sc., Oct. 1992]**

10. The following table gives measurements of blood pressure on subjects by two investigators :

Subject No. :	1	2	3	4	5	6	7	8	9	10
Investigator I :	70	68	56	75	80	90	68	75	56	58
Investigator II :	68	70	52	73	75	78	67	70	54	55

No other details of the experiment were given.

(*i*) If a valid inference has to be drawn about the difference between the investigators, mention the precautions that should have been taken in conducting the experiment with respect to the time of measurement, interval between the first and second measurements, the order in which the investigators measure, etc.

(*ii*) After the experiment was conducted it was discovered that all the subjects were unrelated except that No. 10 was the father of No. 9. Assuming that all the precautions you mention in (*a*) are satisfied, analyse the data to draw an inference on the difference between the investigators. 5 per cent values of the t-statistic corresponding to various degrees of freedom are as follows :

5 per cent values of t...	2·40	2·31	2·26	2·23	2·10	2·09
Degrees of freedom...	7	8	9	10	18	19

11. The following are the values of the cephalic index found in two samples of skulls, one consisting of 15 and the other of 13 individuals.

Sample I :	74·1	77·7	74·0	74·4	73·8	79·3	75·8	82·8
	72·2	75·2	78·2	77·1	78·4	76·3	76·8	
Sample II :	70·8	74·9	74·2	70·4	69·2	72·2	76·8	72·4
	77·4	78·1	72·8	74·3	74·7			

(*i*) Test the hypothesis that the means of population I and population II could be equal.

(*ii*) Is it possible that the sample II has come from a population of mean 72·0 ?

(*iii*) Obtain confidence limits for the mean of population I and for the mean of population II.

(Assume that the distribution of cephallic indices for a homogeneous population is normal.)

12. (*a*) The following table gives the gain in weight in decagrams in a feeding experiment with pigs on the relative value of limestone and bone meal for bone development.

Limestone	49·2	53·3	50·6	52·0	46·8	50·5	52·1	53·0
Bone meal	51·5	54·9	52·2	53·3	51·6	54·1	54·2	53·3

Test for the significance of difference between the means in two ways :

(*i*) by assuming that the values are paired.

(*ii*) by assuming that the values are not paired.

(*b*) The following table shows the mean number of bacterial colonies per plate obtainable by four slightly different methods from soil samples taken at 4 P.M. and 8 P.M. respectively.

Method	*A*	*B*	*C*	*D*
4 P.M.	29·75	27·50	30·25	27·80
8 P.M.	39·20	40·60	36·20	42·40

Are there significantly more bacteria at 8 P.M. than at 4 P.M. ?

[Given $t_{0·05}(3) = 3·18$ and $t_{0·01}(3) = 5·84$]

13. (*a*) It is believed that glucose treatment will extend the sleep time of mice. In an experiment to test this hypothesis ten mice selected at random are given glucose treatment and are found to have a mean hexabarbital sleep time of 47·2 min with a standard deviation of 9·3 min. A further sample of ten untreated mice are found to have a mean hexabarbital sleep time of 28·5 min. with a standard deviation of 7·2 min. Are these results significant evidence in favour of the hypothesis ?

Find 95% confidence limits for the population mean difference in sleep time. State any assumptions made concerning the data in carrying out the test and finding the limits. [*Bangalore Univ. B.E., Oct. 1992*]

(*b*) An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material I were tested, by exposing each piece to a machine measuring wear. Ten pieces of material II were similarly tested. In each case the depth of wear was observed. The sample of material I gave an average (coded) wear 8·5 units with a standard deviation of 0·4 while the sample of material II gave an average of 8·1 and a standard deviation of 0·5. Test the hypothesis that the two types of material exhibit the same mean abrasive wear at the 0·10 level of significance. Assume the populations to be approximately normal with equal variances.

If the level of significance is 0·01, what will be your conclusion ?

[*Delhi Univ. M.E., 1992*]

14·2·11. t-test For Testing Significance of an Observed sample Correlation Coefficient. If r is the observed correlation coefficient in a sample of n pairs of observations from a bivariate normal population, then Prof. Fisher proved that under the null hypothesis $H_0 : \rho = 0$, *i.e., population correlation coefficient is zero,* the statistic :

$$t = \frac{r}{\sqrt{(1 - r^2)}}\sqrt{(n - 2)} \qquad \ldots(14{\cdot}9)$$

follows Student's t-distribution with $(n - 2)$ d.f. (c.f. Remark to § 14·3 page 14·41).

If the value of t comes out to be significant, we reject H_0 at the level of significance adopted and conclude that $\rho \neq 0$, *i.e.*, 'r' is significant of correlation in the population.

If t comes out to be non-significant then H_0 may be accepted and we conclude that variables may be regarded as uncorrelated in the population.

Example 14·12. *A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0·6. Is this significant of correlation in the population ?*

Solution. We set up the null hypothesis, $H_0 : \rho = 0$, *i.e.*, the *observed sample correlation coefficient is not significant of any correlation in the population.*

$$\text{Under } H_0 : \quad t = \frac{r\sqrt{(n - 2)}}{\sqrt{(1 - r^2)}} \sim t_{(n-2)}$$

$$\text{Here} \qquad t = \frac{0{\cdot}6\sqrt{27 - 2}}{\sqrt{(1 - 0{\cdot}36)}} = \frac{3}{\sqrt{0{\cdot}64}} = 3{\cdot}75$$

Tabulated $t_{0.05}$ for $(27 - 2) = 25$ d.f. is 2·06.

Conclusion. Since calculated t is much greater than the tabulated t, it is significant and hence H_0 is discredited at 5% level of significance. Thus we conclude that the variables are correlated in the population.

Example 14·13. *Find the least value of r in a sample of 18 pairs of observations from a bi-variate normal population, significant at 5% level of significance.*

Solution. Here $n = 18$. From the tables $t_{0.05}$ for $(18 - 2) = 16$ d.f. is 2·12

$$\text{Under } H_0 : \rho = 0, \qquad t = \frac{r\sqrt{(n - 2)}}{\sqrt{(1 - r^2)}} \sim t_{(n-2)}$$

In order that the calculated value of t is significant at 5% level of significance, we should have

$$\left|\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}\right| > t_{0\cdot05} \quad \Rightarrow \quad \left|\frac{r\sqrt{16}}{\sqrt{(1-r^2)}}\right| > 2\cdot12$$

$$\Rightarrow \qquad 16r^2 > (2\cdot12)^2(1-r^2) \quad \Rightarrow \quad 20\cdot493r^2 > 4\cdot493$$

$$\Rightarrow \qquad r^2 > \frac{4\cdot493}{20\cdot493} = 0\cdot2192$$

Hence $\qquad |r| > 0\cdot4682$

Example 14·14. *A coefficient of correlation of 0·2 is derived from a random sample of 625 pairs of observations. (i) Is this value of r significant ? (ii) What are the 95% and 99% confidence limits to the correlation coefficient in the population ?*

Solution. *Under the null hypothesis $H_0 : \rho = 0$, i.e., the value of $r = 0\cdot2$ is not significant; the test statistics is :*

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Now
$$t = \frac{0\cdot2 \times \sqrt{(625-2)}}{\sqrt{(1-0\cdot04)}} = 5\cdot09$$

Since $d.f. = 625 - 2 = 623$, the significant values of t are same as in the case of normal distribution, *viz.*, $t_{0\cdot05} = 1\cdot96$ and $t_{0\cdot01} = 2\cdot58$. Since calculated t is much greater than these values; it is highly significant. Hence $H_0 : \rho = 0$ is rejected and we conclude that the sample correlation is significant of correlation in the population.

95% Confidence Limits for ρ (population correlation coefficient) are

$$r \pm 1\cdot96 \text{ S.E. } (r) = r \pm 1\cdot96\,(1-r^2)/\sqrt{n} \qquad \text{[Since } n \text{ large]}$$
$$= 0\cdot2 \pm (1\cdot96 \times 0\cdot96/\sqrt{625}\,)$$
$$= 0\cdot2 \pm 0\cdot075 = (0\cdot125,\ 0\cdot275)$$

99% Confidence Limits for ρ are :

$$0\cdot2 \pm 2\cdot58 \times 0\cdot0384 = 0\cdot2 \pm 0\cdot099 = (0\cdot101,\ 0\cdot299)$$

EXERCISE 14 (d)

1. A restaurant owner ranked his 17 waiters in terms of their speed and efficiency on the job. He correlated these ranks with the total amount of tips each of these waiters received for a one-week period. The obtained value of correlation coefficient is 0·438. What do you conclude ?

Given : $t_{15}\,(0\cdot05) = 2\cdot131$, $t_{16}\,(0\cdot05) = 2\cdot120$ for two-tailed test.

[*Delhi Univ. M.C.A., 1990*]

2. Test the significance of the values of correlation coefficient 'r' obtained from samples of size n pairs from a bivariate normal population.

(*i*) $r = 0\cdot6, n = 38$ $\qquad$ (*ii*) $r = 0\cdot5, n = 11$

Ans. (*i*) $t = 4\cdot5$; Significant at 5% level; $H_0 : \rho = 0$ rejected.

(*ii*) $t = 1\cdot73$; Not significant at 5% level.

(*i*) Consistent Statistic
(*ii*) Unbiased Statistic
(*iii*) Sufficient Statistic
(*iv*) Efficiency. [*Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1982*]

2. What do you understand by Point Estimation ? When would you say that estimate of a parameter is good ? In particular, discuss the requirements of consistency and unbiasedness of an estimate. Give an example to show that a consistent estimate need not be unbiased.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986*]

3. Discuss the terms (*i*) estimate, (*ii*) consistent estimate, (*iii*) unbiased estimate, of a parameter and show that sample mean is both consistent and unbiased estimate of the population mean.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

4. (*a*) If $s_1^2, s_2^2, \ldots, s_r^2$ are r sample variances based on random samples of sizes $n_1, n_2, \ldots, n_r$ respectively, and if T is some statistic given by

$$T = \frac{n_1 s_1^2 + n_2 s_2^2 + \ldots + n_r s_r^2}{a},$$

for estimating σ^2 as an unbiased estimator, find the value, of a, supposing population is very large and for every sample

$$s^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2$$

Ans. $a = (n_1 + n_2 + \ldots + n_r) - r.$

(*b*) If $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_r$ are the sample means based on samples of sizes $n_1, n_2, n_3, \ldots, n_r$ respectively, an unbiased estimator

$$t = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \ldots + n_r \bar{X}_r}{k}$$

has been defined to estimate μ. Find the value of k.

Ans. $k = n_1 + n_2 + \ldots + n_r.$

5. (*a*) For the geometric distribution,

$$f(x, \theta) = \theta\, (1 - \theta)^{x-1}, \ (x = 1, 2, \ldots), 0 < \theta < 1,$$

Obtain an unbiased estimator of $1/\theta$. [**Ans.** $E(\bar{X}) = 1/\theta$.]

(*b*) The random variable X takes the values 1 and 0 with respective probabilities θ and $1 - \theta$. Independent observations $X_1, X_2, \ldots, X_n$ on X are available. Write $\xi = X_1 + X_2 + \ldots + X_n$.

Show that $\xi\, (n - \xi)/n(n - 1)$ is an unbiased estimate of $\theta(1 - \theta)$.

6. Show that if T is an unbiased estimator of a parameter θ, then $\lambda_1 T + \lambda_2$ is an unbiased estimator of $\lambda_1 \theta + \lambda_2$, where λ_1 and λ_2 are known constants, but T^2 is a biased estimator of θ^2.

7. For the following cases determine if the given estimator is unbiased for the parametric function. When it is biased, derive an unbiased estimator from it. $\bar{x}$ is the sample mean.

Proof. Let (x_i, y_i), $(i = 1, 2, ..., n)$ be a random sample of size n drawn from an uncorrelated bivariate normal population $(\rho = 0)$ in which $E(X) = E(Y) = 0$ and $V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$. Let the variable Y be transformed to the variable Z by means of a linear orthogonal transformation, *viz.*,

$$\mathbf{Z} = \mathbf{C}\mathbf{Y}$$

where $\mathbf{Z}_{n \times 1} = (z_1, z_2, \ldots, z_n)'$, $\mathbf{Y}_{n \times 1} = (y_1, y_2, \ldots, y_n)'$ and $C_{n \times n} = (c_{ij})$, C is an orthogonal matrix. Let us, in particular, take

$$c_{11} = c_{12} = \ldots = c_{1n} = 1/\sqrt{n}\ ,$$

so that
$$z_1 = \frac{1}{\sqrt{n}}(y_1 + y_2 + \ldots + y_n) = \sqrt{n}\,\bar{y}$$

Now proceeding as in (*Theorem 13·5*), we get

$$\sum_{i=2}^{n} z_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = ns_Y^2$$

Since in a bivariate normal distribution, the marginal distributions of X and Y are also normal, we have $Y \sim N(0, \sigma_Y^2)$. Hence by Fisher's Lemma (*Theorem 13·4*) z_i, $(i = 1, 2, \ldots, n)$ are independent $N(0, \sigma_Y^2)$.

Now
$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n\, s_X\, s_Y}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, y_i - \bar{y} \sum_{i=1}^{n} (x_i - \bar{x})}{n\, s_X\, s_Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{n s_X\, s_Y}$$

$$\therefore \quad \sqrt{n}\, s_Y\, r = \frac{\Sigma(x_i - \bar{x})\, y_i}{\sqrt{n}\, s_X} = z_2, \text{ (say)}, \qquad \ldots(**)$$

[since the sum of the squares of coefficients of $y_1, y_2, \ldots, y_n$ in (**) is unity.]
From (*) and (**), we get

$$ns_Y^2 = \sum_{i=2}^{n} z_i^2 = \sum_{i=3}^{n} z_i^2 + z_2^2 = \sum_{i=3}^{n} z_i^2 + nr^2 s_Y^2$$

$$\Rightarrow \qquad (1 - r^2)\, n\, s_Y^2 = \sum_{i=3}^{n} z_i^2 \qquad \ldots(***)$$

Since z_i, $(i = 1, 2, \ldots, n)$ are independent $N(0, \sigma_Y^2)$;

$\therefore$ (z_i/σ_Y), $(i = 1, 2, \ldots, n)$ are independent $N(0, 1)$.

Hence from (**),

$$U = \frac{z_2^2}{\sigma_Y^2} = \frac{nr^2 s_Y^2}{\sigma_Y^2}$$

being the square of a standard normal variate is a χ^2-variate with 1 d.f. and from (***),

$$V = \sum_{i=3}^{n} z_i^2/\sigma_Y^2 = \sum_{i=3}^{n} (z_i/\sigma_Y)^2 = \frac{(1-r^2)\, n\, s_Y^2}{\sigma_Y^2},$$

being the sum of squares of $(n-2)$ independent standard normal variates is an independent χ^2-variate with $(n-2)$ d.f.

Further, since z_2 and $(z_3, z_4, \ldots, z_n)$ are independent *r.v.'s*, U and V are independent chi-square variates with 1 and $(n-2)$ d.f. respectively.

$$\therefore \quad \frac{U}{U+V} = \frac{nr^2 s_Y^2/\sigma_Y^2}{[nr^2 s_Y^2 + (1-r^2)\, n\, s_Y^2]/\sigma_Y^2} \sim \beta_1\left(\frac{1}{2}, \frac{n-2}{2}\right)$$

[*c.f.* Theorem 13·2]

$$\Rightarrow \quad r^2 \sim \beta_1\left(\frac{1}{2}, \frac{n-2}{2}\right)$$

Hence the probability function of r^2 is given by

$$dF(r^2) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} (r^2)^{(1/2)-1} [1-r^2]^{\frac{n-2}{2}-1}\, d(r^2),\ 0 \le r^2 \le 1$$

$$\Rightarrow \quad dF(r) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} [1-r^2]^{(n-4)/2}\, dr,\ -1 \le r \le 1$$

the factor 2 disappearing from the fact that total probability in the range $-1 \le r \le 1$ must be unity.

Remark. *If* $\rho = 0$, *then* $t = \dfrac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)}$ *is distributed as Student's* t *with* $(n-2)$ *d. f.*

Proof.
$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \qquad \ldots(*)$$

$$\Rightarrow \quad \frac{t^2}{n-2} = \frac{r^2}{(1-r^2)} = \frac{1}{1-r^2} - 1$$

$$\Rightarrow \quad (1-r^2) = \left[1 + \frac{t^2}{(n-2)}\right]^{-1} \qquad \ldots(**)$$

From (*),

$$dt = \sqrt{(n-2)}\, d[r/\sqrt{(1-r^2)}\,]$$

$$dt = \sqrt{(n-2)} \left[\frac{dr}{\sqrt{(1-r^2)}} + \left(-\frac{r}{2}\right) \frac{(-2r)dr}{(1-r^2)^{3/2}}\right]$$

$$\Rightarrow \quad dt = \sqrt{(n-2)}\, \frac{dr}{\sqrt{(1-r^2)}} \left[1 + \frac{r^2}{1-r^2}\right]$$

$$\Rightarrow \quad dt = \sqrt{(n-2)} \times \frac{dr}{(1-r^2)^{3/2}}, \text{ i.e., } dr = \frac{1}{\sqrt{(n-2)}} (1-r^2)^{3/2}\, dt$$

As r ranges from -1 to 1, from (*), t ranges from $-\infty$ to ∞.

When $\rho = 0$, the p.d.f. of 'r' is given by (14·12) and it transforms to

$$dG(t) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} [1 - r^2]^{(n-4)/2} \frac{1}{\sqrt{(n-2)}} (1 - r^2)^{3/2}\, dt$$

$$= \frac{1}{\sqrt{(n-2)}\, B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{n-2}\right]^{(n-1)/2}},$$

[From (**)]

$$= \frac{1}{\sqrt{(n-2)}\, B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{n-2}\right]^{(n-2+1)/2}},$$

$-\infty < t < \infty$

which is the p.d.f. of t-distribution with $(n-2)$d.f.

Hence $t = \dfrac{r}{\sqrt{(1-r^2)}} \cdot \sqrt{(n-2)} \sim t_{(n-2)}$

Example 14·15. (*a*) *If* (x_i, y_i) *is a random sample drawn from an uncorrelated bivariate normal population, derive the distribution of*

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2\, \Sigma(y_i - \bar{y})^2}}$$

(*b*) *Further, when* $n = 5$ *and if* $P(|r| \geq C) = \alpha$, *show that* C *is a root of the equation,*

$$C\sqrt{(1 - C^2)} + \sin^{-1} C + \frac{\pi(\alpha - 1)}{2} = 0$$

Solution. (*a*) *c.f.* § 14·3.

(*b*) $P(|r| \geq C) = 1 - P(|r| \leq C) = 1 - P(-C \leq r \leq C)$

$$= 1 - 2P(0 \leq r \leq C) = 1 - 2\int_0^C f(r)\, dr$$

[$\because f(r)$ is symmetrical about $r = 0$]

When $n = 5$, $\quad f(r) = \dfrac{1}{B\left(\frac{1}{2}, \frac{3}{2}\right)} \cdot (1 - r^2)^{\frac{1}{2}}\, dr \qquad$ [*c.f.* Equation (14·12)]

$$\therefore \quad P(|r| \geq C) = 1 - 2\frac{\Gamma(2)}{\Gamma(1/2)\Gamma(3/2)} \int_0^C (1 - r^2)^{1/2}\, dr$$

$$= 1 - 2 \times \frac{1}{\frac{1}{2}\pi} \left[\tfrac{1}{2} r (1 - r^2)^{1/2} + \tfrac{1}{2}\sin^{-1} r\right]_0^C$$

$$= 1 - \frac{4}{\pi}\left[\tfrac{1}{2} C (1 - C^2)^{1/2} + \tfrac{1}{2}\sin^{-1} C\right] = \alpha, \text{ (Given)}$$

$$\therefore \qquad 1 - \frac{2}{\pi}\left[C(1-C^2)^{\frac{1}{2}} + \sin^{-1} C\right] = \alpha$$

$$\Rightarrow \qquad C(1-C^2)^{1/2} + \sin^{-1} C + (\alpha - 1)\frac{\pi}{2} = 0$$

14·4. Non-central t-distribution. The non-central t-distribution is the distribution of the ratio of a normal variate with possibly non-zero mean and variance unity, to the square root of an independent χ^2-variate divided by its degrees of freedom. If $X \sim N(\mu, 1)$ and Y is an independent χ^2-variate with n *d.f.*, then

$$t' = \frac{X}{\sqrt{Y/n}} \qquad \ldots(14 \cdot 13)$$

is said to have a non-central t-distribution with n *d.f.* and non-centrality parameter μ. Non-central t-distribution is required for the power functions of certain tests concerning normal population.

p.d.f. of 't''. Since $X \sim N(\mu, 1)$, its *p.d.f.* $f(.)$ is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}(x-\mu)^2\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}(\mu^2 + x^2)\right] \sum_{i=0}^{\infty} \frac{(\mu x)^i}{i\,!}, -\infty < x < \infty$$

Since $Y \sim \chi^2_{(n)}$, its *p.d.f.* $g(.)$ is

$$g(y) = \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-y/2}\, y^{(n/2)-1}, 0 < y < \infty$$

Since X and Y are independent, their joint p.d.f. becomes

$$f(x, y) = \frac{1}{\sqrt{2\pi}\, 2^{n/2}\, \Gamma(n/2)} \exp\left[-\tfrac{1}{2}(\mu^2 + x^2 + y)\right] y^{(n/2)-1} \sum_{i=0}^{\infty} \frac{(\mu x)^i}{i\,!}$$

Let us transform to new variables t' and z by the substitution :

$$t' = \frac{x}{\sqrt{y/n}} = \frac{\sqrt{n}\, x}{\sqrt{y}}, \quad z = +\sqrt{y}$$

$$\Rightarrow \qquad x = zt'/\sqrt{n}, \quad y = z^2$$

Jacobian of transformation J is

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial t'} & \dfrac{\partial x}{\partial z} \\ \dfrac{\partial y}{\partial t'} & \dfrac{\partial y}{\partial z} \end{vmatrix} = \begin{vmatrix} \dfrac{z}{\sqrt{n}} & \dfrac{t'}{\sqrt{n}} \\ 0 & 2z \end{vmatrix} = \frac{2z^2}{\sqrt{n}}$$

The joint p.d.f. of t' and z becomes

$$h(t', z) = \frac{\exp(-\mu^2/2)}{\sqrt{2\pi}\, 2^{n/2}\, \Gamma(n/2)} (z^2)^{\frac{n}{2}-1} \sum_{i=0}^{\infty} \frac{(\mu z t'/\sqrt{n})^i}{i\,!} \cdot \frac{2z^2}{\sqrt{n}};$$

$$\times \exp\left[-\tfrac{1}{2}\left(1+\frac{t'^2}{n}\right)z^2\right]; \qquad -\infty < t' < \infty,\ 0 < z < \infty$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi}\, 2^{(n-1)/2}\, \Gamma(n/2)} \sum_{i=0}^{\infty} \left[\frac{(\mu t')^i}{i!\, n^{(i+1)/2}} \cdot \exp\left\{-\frac{1}{2}\left(1+\frac{t'^2}{n}\right)z^2\right\} z^{n+i}\right]$$

Integrating w.r.t. z in the range 0 to ∞, we get the p.d.f. of t'

$$h_1(t') = \frac{\exp(-\mu^2/2)}{\sqrt{\pi}\, 2^{(n-1)/2}\, \Gamma(n/2)}$$

$$\times \sum_{i=0}^{\infty} \left[\frac{(\mu t')^i}{i!\, n^{(i+1)/2}} \int_0^{\infty} \exp\left\{-\frac{1}{2}\left(1+\frac{t'^2}{n}\right)z^2\right\} z^{n+i}\, dz\right]$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi}\, 2^{(n-1)/2}\, \Gamma(n/2)}$$

$$\times \sum_{i=0}^{\infty} \left[\frac{(\mu t')^i}{i!\, n^{(i+1)/2}} \int_0^{\infty} \exp\left\{-\left(1+\frac{t'^2}{n}\right)v\right\} . (2v)^{(n+i-1)/2}\, dv\right]$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi}\, \Gamma(n/2)} \sum_{i=0}^{\infty} \left[\frac{\mu^i 2^{i/2}\, \Gamma\left(\dfrac{n+i+1}{2}\right)}{i!\, n^{(i+1)/2}} \quad \frac{t'^i}{\left(1+\dfrac{t'^2}{n}\right)^{(n+i+1)/2}}\right]$$

...[14·13(*a*)]

which is the p.d.f. of non-central t-distribution with n d.f. and non-centrality element μ.

Remark. If $\mu = 0$, we get from [14·13 (*a*)]

$$h_1(t') = \frac{1}{\sqrt{\pi}\, \Gamma(n/2)} \cdot \frac{\Gamma[(n+1)/2]}{\sqrt{n}} \cdot \frac{1}{\left[1+\dfrac{t'^2}{n}\right]^{(n+1)/2}}$$

$$= \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \left[1+\frac{t'^2}{n}\right]^{-(n+1)/2}, \quad -\infty < t' < \infty$$

which is the p.d.f. of central t-distribution with n d.f.

14·5. F-statistic. Definition. *If X and Y are two independent chi-square variates with ν_1 and ν_2 d.f. respectively, then F-statistic is defined by*

$$F = \frac{X/\nu_1}{Y/\nu_2} \qquad ...(14·14)$$

In other words, F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F-distribution with (ν_1, ν_2) d.f. with probability function given by

$$f(F)=\frac{\left(\dfrac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{B\left(\dfrac{\nu_1}{2},\dfrac{\nu_2}{2}\right)}\cdot\frac{F^{\frac{\nu_1}{2}-1}}{\left[1+\dfrac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}},\ 0\le F<\infty \qquad \ldots[14\cdot14(a)$$

Remarks 1. The sampling distribution of F-statistic does not involve any population parameters and depends only on the degrees of freedom ν_1 and ν_2.

2. A statistic F following Snedecor's F-distribution with (ν_1, ν_2) *d.f.* will be denoted by $F \sim F(\nu_1, \nu_2)$.

14·5·1 Derivation of Snedecor's F-distribution. Since X and Y are independent chi-square variates with ν_1 and ν_2 d.f. respectively, their joint probability differential is given by

$$dF(x, y) = \left\{\frac{1}{2^{\nu_1/2}\,\Gamma(\nu_1/2)}\exp(-x/2)\,x^{(\nu_1/2)-1}\,dx\right\}$$

$$\times\left\{\frac{1}{2^{\nu_2/2}\,\Gamma(\nu_2/2)}\exp(-y/2)\,y^{(\nu_2/2)-1}\,dy\right\}$$

$$=\frac{1}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}\exp\{-(x+y)/2\}$$

$$\times\, x^{(\nu_1/2)-1}\,y^{(\nu_2/2)-1}\,dx\,dy,\quad 0\le(x,y)<\infty$$

Let us make the following transformation of variables :

$$F=\frac{x/\nu_1}{y/\nu_2}\ \text{ and }\ u=y,\ \text{so that } 0\le F<\infty,\ 0<u<\infty$$

$$\therefore\qquad x=\frac{\nu_1}{\nu_2}Fu=\frac{\nu_1}{\nu_2}Fu\quad\text{and}\quad y=u$$

Jacobian of transformation J is given by

$$J=\frac{\partial(x,y)}{\partial(F,u)}=\begin{vmatrix}\dfrac{\nu_1}{\nu_2}u & 0\\[2ex] \dfrac{\nu_1}{\nu_2}F & 1\end{vmatrix}=\frac{\nu_1 u}{\nu_2}$$

Thus the distribution of the transformed variable is

$$dG(F,u)=\frac{1}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}\exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\}$$

$$\times\left(\frac{\nu_1}{\nu_2}Fu\right)^{(\nu_1/2)-1}u^{(\nu_2/2)-1}\,|J|\,du\,dF$$

$$=\frac{(\nu_1/\nu_2)^{\nu_1/2}}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}\exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\}$$

$$\times\, u^{\{(\nu_1+\nu_2)/2\}-1}\,F^{(\nu_1/2)-1}\,du\,dF;\ 0<u<\infty,\ 0\le F<\infty$$

Integrating out u over the range 0 to ∞, the distribution of F becomes

$$g_1(F)\,dF = \frac{(\nu_1/\nu_2)^{(\nu_1/2)} F^{(\nu_1/2)-1}\,dF}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \times \left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\} u^{\{(\nu_1+\nu_2)/2\}-1}\,du\right]$$

$$= \frac{(\nu_1/\nu_2)^{(\nu_1/2)} F^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \times \frac{\Gamma[(\nu_1+\nu_2)/2]}{\left[\frac{1}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right]^{(\nu_1+\nu_2)/2}}\,dF$$

$$\therefore \quad g_1(F) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left[1+\frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}},\ 0 \le F < \infty$$

which is the required probability function of F-distribution with (ν_1, ν_2) d.f.

Aliter $$F = \frac{x/\nu_1}{y/\nu_2}$$

$\therefore \quad \dfrac{\nu_1}{\nu_2}F = \dfrac{x}{y}$, being the ratio of two independent chi-square variates with ν_1 and ν_2 d.f. respectively is a $\beta_2\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)$ variate. Hence the probability function of F is given by

$$dP(F) = \frac{1}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \cdot \frac{\left(\frac{\nu_1}{\nu_2}F\right)^{(\nu_1/2)-1}}{\left[1+\frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}}\,d\left(\frac{\nu_1}{\nu_2}F\right)$$

$$= \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left[1+\frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}}\,dF,\ 0 \le F < \infty$$

14·5·2. Constants of F-distribution.

$$\mu'_r \text{ (about origin)} = E(F^r) = \int_0^\infty F^r f(F)\,dF$$

$$= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \int_0^\infty F^r \frac{F^{(\nu_1/2)-1}}{\left[1+\frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}}\,dF \qquad \ldots(*)$$

To evaluate the integral, put

$$\frac{\nu_1}{\nu_2}F = y, \text{ so that } dF = \frac{\nu_2}{\nu_1}dy$$

$$\therefore \quad \mu_r' = \frac{[\nu_1/\nu_2]^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{\left(\frac{\nu_2}{\nu_1}\, y\right)^{r + (\nu_1/2) - 1}}{[1 + y]^{(\nu_1 + \nu_2)/2}} \left(\frac{\nu_2}{\nu_1}\right) dy$$

$$= \frac{\left(\frac{\nu_2}{\nu_1}\right)^r}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{y^{r + (\nu_1/2) - 1}}{[1 + y]^{(\nu_1/2) + r + \{(\nu_2/2) - r\}}}\, dy$$

$$= \left(\frac{\nu_2}{\nu_1}\right)^r \cdot \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot B\left(r + \frac{\nu_1}{2}, \frac{\nu_2}{2} - r\right), \nu_2 > 2r \qquad \ldots(14\cdot15)$$

Aliter for (14·15). (14·15) could also be obtained by substituting $\frac{\nu_1}{\nu_2} F = \tan^2 \theta$ in (*) and using the Beta integral :

$$2 \int_0^{\pi/2} \sin^p \theta \cos^q \theta \, d\theta = B\left(\frac{p + 1}{2}, \frac{q + 1}{2}\right)$$

$$\therefore \quad \mu_r' = \left(\frac{\nu_2}{\nu_1}\right)^r \cdot \frac{\Gamma[r + (\nu_1/2)]\, \Gamma[(\nu_2/2) - r]}{\Gamma(\nu_1/2)\, \Gamma(\nu_2/2)}, \; r < \frac{\nu_2}{2}. \qquad \ldots(14\cdot16)$$

In particular

$$\mu_1' = \frac{\nu_2}{\nu_1} \cdot \frac{\Gamma[1 + (\nu_1/2)]\, \Gamma[(\nu_2/2) - 1]}{\Gamma(\nu_1/2)\, \Gamma(\nu_2/2)}$$

$$= \frac{\nu_2}{\nu_2 - 2}, \nu_2 > 2 \quad [\because \Gamma(r) = (r - 1)\, \Gamma(r - 1)] \qquad \ldots[14\cdot16(a)]$$

Thus the mean of F-distribution is independent of ν_1.

$$\mu_2' = \left(\frac{\nu_2}{\nu_1}\right)^2 \cdot \frac{\Gamma[(\nu_1/2) + 2]\, \Gamma[(\nu_2/2) - 2]}{\Gamma(\nu_1/2)\, \Gamma(\nu_2/2)}$$

$$= \left(\frac{\nu_2}{\nu_1}\right)^2 \cdot \frac{[(\nu_1/2) + 1]\, (\nu_1/2)}{[(\nu_2/2) - 1]\, [(\nu_2/2) - 2]}$$

$$= \frac{\nu_2^2(\nu_1 + 2)}{\nu_1(\nu_2 - 2)\, (\nu_2 - 4)}, \nu_2 > 4.$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = \frac{\nu_2^2(\nu_1 + 2)}{\nu_1(\nu_2 - 2)(\nu_2 - 4)} - \frac{\nu_2^2}{(\nu_2 - 2)^2}$$

$$= \frac{2\nu_2^2\, (\nu_2 + \nu_1 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, \nu_2 > 4 \qquad \ldots[14\cdot16(b)]$$

Similarly, on putting $r = 3$ and 4 in μ_r', we get μ_3' and μ_4' respectively, from which the central moments μ_3 and μ_4 can be obtained.

Remark. It has been proved that for large degrees of freedom, ν_1 and ν_2, F tends to $N[1, 2\{(1/\nu_1) + (1/\nu_2)\}]$ variate.

14·5·3. Mode and Points of Inflexion of F-distribution. We have

$$\log f(F) = C + \{(\nu_1/2) - 1\} \log F - \left(\frac{\nu_1 + \nu_2}{2}\right) \log \{1 + (\nu_1/\nu_2) F\}$$

where C is a constant independent of F.

$$\frac{\partial}{\partial F} [\log f(F)] = \left(\frac{\nu_1}{2} - 1\right) \cdot \frac{1}{F} - \frac{(\nu_1 + \nu_2)}{2} \cdot \frac{1}{\left[1 + \frac{\nu_1}{\nu_2} F\right]} \cdot \frac{\nu_1}{\nu_2}$$

$$f'(F) = \frac{\partial}{\partial F} f(F) = 0 \quad \Rightarrow \quad \frac{\nu_1 - 2}{2F} - \frac{\nu_1 (\nu_1 + \nu_2)}{2(\nu_2 + \nu_1 F)} = 0$$

Hence $$F = \frac{\nu_2 (\nu_1 - 2)}{\nu_1 (\nu_2 + 2)} \qquad ...(14·17)$$

It can be easily verified that at this point $f''(F) < 0$. Hence

$$\text{Mode} = \frac{\nu_2 (\nu_1 - 2)}{\nu_1 (\nu_2 + 2)}$$

Remarks 1. Since $F > 0$, mode exists if and only if $\nu_1 > 2$.

2. $$\text{Mode} = \left(\frac{\nu_2}{\nu_2 + 2}\right) \cdot \left(\frac{\nu_1 - 2}{\nu_1}\right)$$

Hence mode of F-distribution is always less than unity.

3. The points of inflexion of F-distribution exist for $\nu_1 > 4$ and are equidistant from mode.

Proof. We have $$\frac{\nu_1}{\nu_2} F = \frac{X}{Y} \sim \beta_2 (l, m), \qquad (*)$$

where $l = \nu_1/2$ and $m = \nu_2/2$. We now find the points of inflexion of Beta distribution of second kind with parameters l and m.

If $X \sim \beta_2 (l, m)$, its *p.d.f.* is

$$f(x) = \frac{1}{\beta(l, m)} \cdot \frac{x^{l-1}}{(1 + x)^{l+m}} ; 0 \le x < \infty \qquad ...(**)$$

Points of inflexion are the solution of

$$f''(x) = 0 \text{ and } f'''(x) \neq 0$$

From (**),

$$\log f(x) = -\log \beta(l, m) + (l - 1) \log x - (l + m) \log (1 + x)$$

Differentiating twice w.r.t x, we get

$$\frac{f'(x)}{f(x)} = \frac{l - 1}{x} - \frac{l + m}{1 + x} \qquad ...(***)$$

$$\frac{f(x) f''(x) - [f'(x)]^2}{[f(x)]^2} = -\left(\frac{l - 1}{x^2}\right) + \frac{l + m}{(1 + x)^2}$$

If $f''(x) = 0$, then we get

$$-\left[\frac{f'(x)}{f(x)}\right]^2 = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2}$$

$$\Rightarrow \quad -\left[\frac{l-1}{x} - \frac{l+m}{1+x}\right]^2 = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2} \qquad \text{[On using (***)]}$$

$$\Rightarrow \quad \frac{l-1}{x^2}(l-1-1) - 2\frac{(l-1)(l+m)}{x(1+x)} + \frac{l+m}{(1+x)^2} \times (l+m+1) = 0$$

$$\Rightarrow (l-1)(l-2)(1+x)^2 - 2x(1+x)(l-1)(l+m) + x^2(l+m)(l+m+1) = 0 \qquad \ldots(****)$$

which is a quadratic in x. It can be easily verified that at these values of x, $f'''(x) \neq 0$, if $l > 2$.

The roots of (****) give the points of inflexion of $\beta_2(l, m)$ distribution. The sum of the points of inflexion is equal to the sum of roots of (****) and is given by :

$$-\left[\frac{\text{Coefficient of } x \text{ in (****)}}{\text{Coefficient of } x^2 \text{ in (****)}}\right]$$

$$= -\left[\frac{2(l-1)(l-2) - 2(l-1)(l+m)}{(l-1)(l-2) - 2(l-1)(l+m) + (l+m)(l+m+1)}\right]$$

$$= \frac{2(l-1)[(l+m)-(l-2)]}{(l-1)(l-2) - (l-1)(l+m) - (l-1)(l+m) + (l+m)(l+m+1)}$$

$$= \frac{2(l-1)(m+2)}{(l-1)[(l-2-l-m] + (l+m)[l+m+1-l+1)}$$

$$= \frac{2(l-1)(m+2)}{-(l-1)(m+2) + (l+m)(m+2)}$$

$$= \frac{2(l-1)}{l+m-l+1} = \frac{2(l-1)}{(m+1)}$$

$\therefore$ Sum of points of inflexion of $\left(\frac{\nu_1}{\nu_2}F\right)$ distribution

$$= \frac{2(l-1)}{(m+1)} = \frac{2\left(\frac{\nu_1}{2}-1\right)}{\left(\frac{\nu_2}{2}+1\right)} = \frac{2(\nu_1-2)}{(\nu_2+2)}.$$

$\Rightarrow$ Sum of points of inflexion of $F(\nu_1, \nu_2)$ distribution

$$= \frac{\nu_2}{\nu_1} \cdot \frac{2(\nu_1-2)}{(\nu_2+2)}, \text{ provided } l = \frac{\nu_1}{2} > 2$$

$$= 2\frac{\nu_2(\nu_1-2)}{\nu_1(\nu_2+2)}$$

$= 2$ Mode, provided $\nu_1 > 4$

Hence the points of inflexion of $F(\nu_1, \nu_2)$ distribution, when they exist, (*i.e.*, when $\nu_1 > 4$), are equidistant from the mode.

4. Karl Pearson's coefficient of skewness is given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma} > 0,$$

since mean > 1 and mode < 1. Hence F-distribution is highly positively skewed.

5. The probability $p(F)$ increases steadily at first until it reaches its peak (corresponding to the modal value which is less than 1) and then decreases slowly so as to become tangential at $F = \infty$, *i.e.*, F-axis is an asymptote to the right tail.

Example 14·16. *When $\nu_1 = 2$, show that the singificance level of F corresponding to a significant probability p is*

$$F = \frac{\nu_2}{2}\left(p^{-(2/\nu_2)} - 1\right)$$

where ν_1 an ν_2 have their usual meanings.

Solution. When $\nu_1 = 2$,

$$dP(F) = \frac{1}{B\left(1, \frac{\nu_2}{2}\right)} \cdot \frac{2}{\nu_2} \cdot \frac{dF}{\left[1 + \frac{2}{\nu_2} F\right]^{(\nu_2/2)+1}} \qquad \text{(c.f. §14·14}a\text{)}$$

$$= \frac{\Gamma(\frac{\nu_2}{2} + 1)}{\Gamma(1)\Gamma(\nu_2/2)} \times \frac{\frac{2}{\nu_2}}{\left(\frac{2}{\nu_2}\right)^{(\nu_2/2)+1}\left[F + \frac{\nu_2}{2}\right]^{(\nu_2/2)+1}} dF$$

$$= \frac{\left[\frac{\nu_2}{2}\right]^{(\nu_2/2)+1}}{\left[F + \frac{\nu_2}{2}\right]^{(\nu_2/2)+1}} dF \cdot$$

Hence
$$p = \int_F^\infty f(F)\, dF$$

$$= \left[\frac{\nu_2}{2}\right]^{(\nu_2/2)+1} \times \int_F^\infty \frac{dF}{\left[F + \frac{\nu_2}{2}\right]^{(\nu_2/2)+1}}$$

$$= \left[\frac{\nu_2}{2}\right]^{(\nu_2/2)+1} \times \left| \frac{\left[F + \frac{\nu_2}{2}\right]^{-(\nu_2/2)}}{-\frac{\nu_2}{2}} \right|_F^\infty$$

$$= \left[\frac{\left(\frac{\nu_2}{2}\right)}{F + \frac{\nu_2}{2}}\right]^{\nu_2/2} = \frac{1}{\left[1 + \frac{2}{\nu_2} F\right]^{\nu_2/2}}$$

$$\Rightarrow \quad p^{-(2/\nu_2)} = 1 + \frac{2F}{\nu_2} \quad \Rightarrow \quad F = \frac{\nu_2}{2}\left[p^{-(2/\nu_2)} - 1\right]$$

Example 14·17. *If $F(n_1, n_2)$ represent an F-variate with n_1 and n_2 d.f., prove that $F(n_2, n_1)$ is distributed as $1/F(n_1, n_2)$ variate. Deduce that*

$$P[F(n_1, n_2) \geq c] = P\left[F(n_2, n_1) \leq \frac{1}{c}\right]$$

Or

Show how the probability points of $F(n_2, n_1)$ can be obtained from those of $F(n_1, n_2)$.

Solution. Let X and Y be independent chi-square variates with n_1 and n_2 d.f. respectively. Then by definition, we have

$$F = \frac{(X/n_1)}{(Y/n_2)} \sim F(n_1, n_2)$$

$$\therefore \quad \frac{1}{F} = \frac{(Y/n_2)}{(X/n_1)} \sim F(n_2, n_1) \qquad \ldots(*)$$

Hence the result.

We have :

$$P[F(n_1, n_2) \geq c] = P\left[\frac{1}{F(n_1, n_2)} \leq \frac{1}{c}\right]$$

$$= P\left[F(n_2, n_1) \leq \frac{1}{c}\right] \qquad \text{[From (*)]}$$

Remark. $P[F(n_1, n_2) = c] = P\left[F(n_2, n_1) = \dfrac{1}{c}\right]$

Let $\quad P[F(n_1, n_2) \geq c] = \alpha$

i.e., let c be the upper α-significant point of $F(n_1, n_2)$ distribution.

$$\therefore \quad 1 - \alpha = 1 - P[F(n_1, n_2) \geq c] = 1 - P\left[\frac{1}{F(n_1, n_2)} \leq \frac{1}{c}\right]$$

$$\Rightarrow \quad \alpha = P\left[F(n_2, n_1) \leq \frac{1}{c}\right] = 1 - P\left[F(n_2, n_1) \geq \frac{1}{c}\right]$$

$$\Rightarrow \quad P\left[F(n_2, n_1) \geq \frac{1}{c}\right] = 1 - \alpha$$

Thus $(1 - \alpha)$ significant points of $F(n_2, n_1)$ distribution are the reciprocal of α-significant points of $F(n_1, n_2)$ distribution, *e.g.*,

$$F_{8,4}(0{\cdot}05) = 6{\cdot}04 \Rightarrow F_{4,8}(0{\cdot}95) = \frac{1}{6{\cdot}04}$$

Example 14·18. *Prove that if $n_1 = n_2$, the median of F-distribution is at $F = 1$ and that the quartiles Q_1 and Q_3 satisfy the condition $Q_1 Q_3 = 1$.*

[Delhi Univ. B.Sc. (Stat.Hons.), 1989]

Solution. Since $n_1 = n_2 = n$, (say), the median (M) of $F(n_1, n_2) = F(n, n)$ distribution is given by :

$$P[F(n, n) \le M] = 0.5 \qquad \ldots(*)$$

$$\Rightarrow \quad P\left[\frac{1}{F(n, n)} \ge \frac{1}{M}\right] = 0.5$$

$$\Rightarrow \quad P\left[F(n, n) \ge \frac{1}{M}\right] = 0.5 \qquad \left[\because \frac{1}{F(m, n)} = F(n, m)\right]$$

$$\Rightarrow \quad P\left[F(n, n) \le \frac{1}{M}\right] = 1 - P\left[F(n, n) \ge \frac{1}{M}\right]$$

$$= 1 - 0.5$$

$$= 0.5 \qquad \ldots(**)$$

From (*) and (**), we get

$$M = \frac{1}{M} \quad \Rightarrow \quad M^2 = 1 \quad \Rightarrow \quad M = 1$$

the negative value $M = -1$, is discarded since $F > 0$.

Hence the median of $F(n, n)$ distribution is at $F = 1$.

Similarly, by definition of Q_1 and Q_3, we have :

$$P[F(n, n) \le Q_1] = 0.25 \qquad \ldots(****)$$

and $$P[F(n, n) \ge Q_3] = 0.25$$

$$\Rightarrow \quad P\left[\frac{1}{F(n, n)} \le \frac{1}{Q_3}\right] = 0.25$$

$$\Rightarrow \quad P\left[F(n, n) \le \frac{1}{Q_3}\right] = 0.25 \qquad \left[\because \frac{1}{F(m, n)} = F(n, m)\right] \qquad \ldots(***)$$

From (***) and (****), we get

$$Q_1 = \frac{1}{Q_3} \quad \Rightarrow \quad Q_1 Q_3 = 1$$

Example 14·19. *Let $X_1 X_2, \ldots, X_n$ be a random sample from $N(0, 1)$.*

Define $$\bar{X}_k = \frac{1}{k}\sum_1^k X_i \quad \text{and} \quad \bar{X}_{n-k} = \frac{1}{n-k}\sum_{k+1}^{n} X_i$$

Find the distribution of :

(*a*) $\frac{1}{2}\left(\bar{X}_k + \bar{X}_{n-k}\right)$, (*b*) $k\bar{X}_k^2 + (n-k)\,\bar{X}^2_{n-k}$

(*c*) X_1^2/X_2^2, (*d*) X_1/X_2

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989*]

Solution. (*a*) Since $X_1, X_2, \ldots, X_n$ is a random sample from $N(0, 1)$,

$$\bar{X}_k \sim N\left(0, \frac{1}{k}\right) \quad \text{and} \quad \bar{X}_{n-k} \sim N\left(0, \frac{1}{n-k}\right) \qquad \ldots(*)$$

Further, since $(X_1, X_2, \ldots, X_k)$ and $(X_{k+1}, X_{k+2}, \ldots, X_n)$ are independent, $\bar{X}_k$ and $\bar{X}_{n-k}$ are independent. Hence,

$$\tfrac{1}{2}(\bar{X}_k + \bar{X}_{n-k}) = \tfrac{1}{2}\bar{X}_k + \tfrac{1}{2}\bar{X}_{n-k} \sim N\left(0, \frac{1}{4k} + \frac{1}{4(n-k)}\right)$$

$$\Rightarrow \qquad \tfrac{1}{2}(\bar{X}_k + \bar{X}_{n-k}) \sim N\left(0, \frac{n}{4k\,(n-k)}\right)$$

(*b*) From (*), we get

$$\frac{\bar{X}_k}{\sqrt{1/k}} \sim N(0, 1) \quad \text{and} \quad \frac{\bar{X}_{n-k}}{\sqrt{1/(n-k)}} \sim N(0, 1)$$

$$\Rightarrow \qquad k\,\bar{X}_k^2 \sim \chi^2_{(1)} \quad \text{and} \quad (n-k)\,\bar{X}^2_{n-k} \sim \chi^2_{(1)}$$

Since $\bar{X}_k$ and $\bar{X}_{n-k}$ are independent, by additive property of chi-square distribution,

$$k\,\bar{X}_k^2 + (n-k)\,\bar{X}^2_{n-k} \sim \chi^2_{(1+1)} = \chi^2_{(2)}$$

(*c*) Since $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$ are independent,

$$X_1^2 \sim \chi^2_{(1)} \text{ and } X_2^2 \sim \chi^2_{(1)},$$

are also independent. Hence by definition of F-statistic,

$$\frac{X_1^2/1}{X_2^2/1} \sim F_{(1,1)} \Rightarrow \frac{X_1^2}{X_2^2} \sim F_{(1,1)}$$

(*d*) X_1/X_2, being the ratio of two independent standard normal variates is a standard Cauchy variate. [See Example 8·43].

EXERCISE 14(e)

1. (*a*) Derive the distribution of $F = S_1^2/S_2^2$, where S_1^2 and S_2^2 are two independent unbiased estimates of the common population variance σ^2, defined by

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2; \; S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2$$

(*b*) Find the limiting form when the degrees of freedom of the χ^2 in the denominator tend to infinity and give an intuitive justification of the result.

2. (*a*) If $X_1, X_2, \ldots, X_m, X_{m+1}, \ldots\ldots, X_{m+n}$ are independent normal variates with zero mean and standard deviation σ, obtain the distribution of

$$\sum_{i=1}^{m} X_i^2 \Big/ \sum_{i=m+1}^{m+n} X_i^2$$

Ans. $F(m, n)$.

(*b*) If X has an F distribution with n_1 and n_2 d.f., find the distribution of $1/X$ and give one use of this result.

(*c*) If X is t-distributed, show that X^2 is F-distributed.

[***Delhi Univ. B.Sc. (Maths. Hons.), 1990***]

Hint. See § 14·5·6.

3. (a) Derive the distribution of the F-statistic on (n_1, n_2) degrees of freedom and show that the statistic $\left(1 + \frac{n_1}{n_2} F\right)^{-1}$ has a Beta distribution.

(b) Show that the probability curve of the distribution of F is positively skewed.

4. Prove the following :

(i) $$F_{n_1, n_2} = \frac{1}{F_{n_2, n_1}}$$

(ii) $$F_{n_1, n_2} = \frac{n_2}{n_1} \cdot \frac{x}{1 - x},$$ where x has Beta-distribution.

5. If X and Y are independent chi-square variates with v_1 and v_2 d.f. respectively, show that $U = X + Y$ and $V = \frac{v_2 X}{v_1 Y}$ are independently distributed. Find the distribution of V.

6. Prove that if X has the F-distribution with (m, n) d.f. and Y has the F-distribution with (n, m) d.f., then for every $a > 0$,

$$P(X \leq a) + P\left\{Y \leq \frac{1}{a}\right\} = 1$$

7. Show that the mode of the F-distribution with v_1 (≥ 2), v_2 d.f. is given by $\frac{v_2 (v_1 - 2)}{v_1 (v_2 + 2)}$ and is always less than unity.

8. X is F-variate with 2 and n $(n \geq 2)$ degrees of freedom. Show that

$$P(F \geq k) = \left(1 + \frac{2k}{n}\right)^{-n/2}$$

[*Gujarat Univ. B.Sc., 1992*]

Deduce the significance level of F corresponding to the significance level of probability P.

9. Let X_1, X_2 be independent random variables following the density law $f(x) = e^{-x}$, $0 < x < \infty$. Show that

$Z = X_1/X_2$, has an F-distribution.

10. (a) If $X \sim F(n_1, n_2)$, show that its mean is independent of n_1.

(b) Obtain the mode of F-distribution with (n_1, n_2) d.f. and show that it lies between 0 and 1.

(c) Show that for F-distribution with n_1 and n_2 d.f., the points of inflexion exist if $n_1 > 4$ and are equidistant from the mode.

11. X is a binomial variate with parameters n and p and F_{v_1, v_2} is an F-statistic with v_1 and v_2 d.f. Prove that

$$P(X \leq k - 1) = P\left[F_{2k,\, 2(n-k+1)} > \frac{n - k + 1}{k} \cdot \frac{p}{1 - p}\right]$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1985*]

Hint. If $X \sim B(n, p)$, then we have [c.f. Example 7·23]

$$P(X \le k-1) = (n-k+1).\binom{n}{k-1}\int_0^q t^{n-k}(1-t)^{k-1}\,dt$$

$$= \frac{1}{B(k,\, n-k+1)}\int_0^q t^{n-k}(1-t)^{k-1}\,dt$$

$$P\left[F_{2k,\,2(n-k+1)} > \frac{n-k+1}{k}\left(\frac{p}{1-p}\right)\right] = \int_{\frac{n-k+1}{k}\cdot\frac{p}{q}}^{\infty} p[F_{2k,\,2(n-k+1)}]\,dF$$

$$= \frac{1}{B(k,\, n-k+1)}\int_{\frac{n-k+1}{k}\cdot\frac{p}{q}}^{\infty} \frac{[k/(n-k+1)]^k \cdot F^{k-1}\,dF}{\left[1+\dfrac{kF}{n-k+1}\right]^{n+1}}$$

$$= \frac{1}{B(k,\, n-k+1)}\int_0^q y^{n-k}(1-y)^{k-1}\,dy$$

where $\quad 1 + \dfrac{kF}{n-k+1} = \dfrac{1}{y}.$

12. (a) If $X \sim F(n_1, n_2)$ distribution, show that

$$U = \frac{n_1 X}{n_2 + n_1 X} \sim \beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$$

[***Delhi Univ. B.Sc. (Maths. Hons.), 1992***]

Hence obtain the distribution function of X.

Hint. The distribution function of $X \sim F(n_1, n_2)$ is given by

$$G_X(x) = \int_0^x f(F)\,dF = \int_0^y h(u)du, \quad \left[y = \frac{n_1 x}{n_2 + n_1 x}\right]$$

$$= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)}\int_0^y u^{\frac{n_1}{2}-1}(1-u)^{\frac{n_2}{2}-1}\,du$$

$$= I_y\left(\frac{n_1}{2}, \frac{n_2}{2}\right),$$

where $\quad I_x(p, q) = \dfrac{1}{B(p,q)}\displaystyle\int_0^x t^{p-1}(1-t)^{q-1}\,dt,$

is the incomplete Beta function. Hence the distribution function of F distribution can be obtained from the tables of incomplete Beta function.

(b) $X \sim F(m, n)$, show that

$$W = \frac{m X/n}{1 + (m X/n)} \sim \beta_1\left(\tfrac{1}{2}m,\ \tfrac{1}{2}n\right)$$

Deduce the variance of X from p.d.f. of W.

[***Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989***]

13. Let X_1 and X_2 be a random sample of size 2 form $N(0, 1)$ and Y_1 and Y_2 be a random sample of size 2 from $N(1, 1)$, and let the Y_i's be independent of the X_i's. Find the distribution of the following :

(i) $(X_1 - X_2)/\sqrt{2}$ (ii) $(X_1 + X_2)^2/(X_2 - X_1)^2$

(iii) $\bar{X} + \bar{Y}$ (iv) $(Y_1 + Y_2 - 2)^2/(X_2 - X_1)^2$

(v) $(X_1 + X_2)/\sqrt{[(X_2 - X_1)^2 + (Y_2 - Y_1)^2]/2}$

(vi) $[(Y_1 - Y_2)^2 + (X_1 - X_2)^2 + (X_1 + X_2)^2]/2$

[*Delhi Univ. B.Sc. (Maths. Hons.), 1988, 1987*]

Ans. (i) $N(0, 1)$, (ii) $F(1, 1)$, (iii) $N(1, 1)$
(iv) $F(1, 1)$, (v) $t_{(2)}$, (vi) $\chi^2_{(3)}$.

14. Let $X_i \sim N(i, i^2), i = 1, 2, 3$ be independent random variables. Using only the three random variables X_1, X_2, and X_3, give an example of a statistic that has :

(i) A chi-square distribution with 3 d.f.
(ii) An F-distribution with (1, 2) d.f.
(iii) A t-distribution with 2 d.f.

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1986*]

Ans. **Hint.** $Z_i = (X_i - i)/i,\ i = 1, 2, 3,$ are i.i.d. $N(0, 1)$.

(i) $\sum_{i=1}^{3} Z_i^2 \sim \chi^2_{(3)}$; (ii) $\dfrac{2Z_1^2}{Z_2^2 + Z_3^2} \sim F(1, 2)$; (iii) $\dfrac{Z_1}{[(Z_2^2 + Z_3^2)/2]^{1/2}} \sim t_{(2)}$

15. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Define :

$$\bar{X}_k = \frac{1}{k}\sum_{1}^{k} X_i,\quad \bar{X}_{n-k} = \frac{1}{n-k}\sum_{k+1}^{n} X_i,\quad \bar{X} = \frac{1}{n}\sum_{1}^{n} X_i$$

$$S_k^2 = \frac{1}{k-1}\sum_{1}^{k} X_i - \bar{X}_k)^2,\quad S^2_{n-k} = \frac{1}{n-k-1}\sum_{k+1}^{n}(X_i - \bar{X}_{n-k})^2$$

and $$S^2 = \frac{1}{n-1}\sum_{1}^{n}(X_i - \bar{X})^2.$$

Answer the following questions :

(i) What is the distribution of

$$\sigma^{-2}[(k-1)S^2_k + (n-k-1)S^2_{n-k}]\ ?$$

(ii) What is the distribution of S^2_k/S^2_{n-k} ?

(iii) What is the distribution of $(\bar{X} - \mu)\sqrt{n}/S$?

[*Delhi Univ. B.Sc. (Maths Hons.), 1989*]

(iv) What is the distribution of $\frac{1}{2}(\bar{X}_k + \bar{X}_{n-k})$?

(v) What is the distribution of $(X_i - \mu)^2/\sigma^2$?

Ans. (i) $\chi^2_{(k-1)+(n-k-1)} = \chi^2_{(n-2)}$; (ii) $F_{(k-1,\, n-k-1)}$

(iii) $t_{(n-1)}$; (iv) $N\left(\mu, \dfrac{n\sigma^2}{4k(n-k)}\right)$, (v) $\chi^2_{(1)}$

16. If $X \sim F(1, n)$, show that

$$\left(n - \frac{1}{2}\right) \log\ [1 + (X/n)] \sim \chi^2_{(1)},$$

for large n.

17. If X_1, X_2, X_3 and X_4 are independent observations from N (0, 1) population, state giving reasons, the sampling distributions of

$$(i)\ U = \frac{\sqrt{2}\, X_3}{\sqrt{X_1^2 + X_2^2}} \quad \text{and} \quad (ii)\ V = \frac{3X_4^2}{X_1^2 + X_2^2 + X_3^2}$$

Ans. (*i*) $U \sim t_{(2)}$; (*ii*) $V \sim F(1, 3)$.

18. Let (X_1, X_2) be a random sample from $N(0, 1)$. Answer the following, giving reasons :

(*i*) What is the distribution of $(X_2 - X_1)^2/2$?

(*ii*) What is the distribution of $(X_1 + X_2)^2/(X_2 - X_1)^2$?

(*iii*) What is the distribution of $(X_2 + X_1)/\sqrt{(X_1 - X_2)^2}$?

(*iv*) What is the distribution of $1/Z$, if $Z = X_1^2/X_2^2$?

[*Delhi Univ. B.Sc. (Maths. Hons.), 1992*]

Ans. (*i*) $\chi^2_{(1)}$; (*ii*) $F(1, 1)$; (*iii*) Standard Cauchy ; (*iv*) $F(1, 1)$

14·5·4. Applications of F-distribution. F-distribution has the following applications in Statistical theory.

14·5·5. F-test for Equality of Population Variances. Suppose we want to test (*i*) whether two independent samples x_i, $(i = 1, 2, \ldots, n_1)$ and y_j, $(j = 1, 2, \ldots, n_2)$ have been drawn from the normal populations with the same variance σ^2, (say), or (*ii*) whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis (H_0) that (*i*) $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, *i.e., the population variances are equal or (ii) Two independent estimates of the population variance are homogeneous*, the statistic F is given by

$$F = \frac{S_X^2}{S_Y^2} \qquad \ldots(14\cdot18)$$

where $S_X^2 = \dfrac{1}{n_1 - 1} \displaystyle\sum_{i=1}^{n_1} (x_i - \bar{x})^2$ and $S_Y^2 = \dfrac{1}{n_2 - 1} \displaystyle\sum_{j=1}^{n_2} (y_j - \bar{y})^2$ ...(14·18*a*)

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows Snedecor's F-distribution with (ν_1, ν_2) d.f. [where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$].

Proof.

$$F = \frac{S_X^2}{S_Y^2} = \left[\frac{n_1}{n_1 - 1} s_X^2\right] \Bigg/ \left[\frac{n_2}{n_2 - 1} \cdot s_Y^2\right]$$

$$= \left[\frac{n_1 s_X^2}{\sigma_X^2} \cdot \frac{1}{(n_1 - 1)}\right] \Bigg/ \left[\frac{n_2 s_Y^2}{\sigma_Y^2} \cdot \frac{1}{(n_2 - 1)}\right]$$

$(\because \sigma_X^2 = \sigma_Y^2 = \sigma^2$ under $H_0)$

Since $\dfrac{n_1 s_X^2}{\sigma_X^2}$ and $\dfrac{n_2 s_Y^2}{\sigma_Y^2}$ are independent chi-square variates with $(n_1 - 1)$ and $(n_2 - 1)$ d.f. respectively, F follows Snedecor's F-distribution with $(n_1 - 1, n_2 - 1)$ d.f. (c.f. § 14·5).

Remarks 1. In (14·18), greater of the two variances S_X^2 and S_Y^2 is to be taken in the numerator and n_1 corresponds to the greater variance.

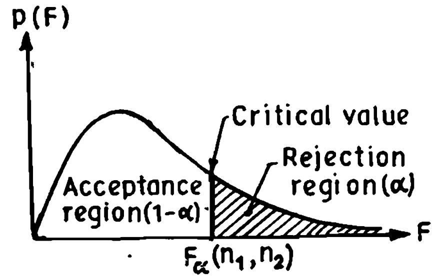
By comparing the calculated value of F obtained by using (14·18) for the two given samples with the tabulated value of F for (n_1, n_2) d.f. at certain level of significance (5% or 1%), H_0 is either rejected or accepted.

2. *Critical values of F-distribution.* The available F-tables (given in the Appendix at the end of the book) give the critical values of F for the right-tailed test, *i.e.*, the critical region is determined by the right-tail areas. Thus the significant value $F_\alpha(n_1, n_2)$ at level of significance α and (n_1, n_2) *d.f.* is determined by

$$P[F > F_\alpha(n_1, n_2)] = \alpha, \qquad \ldots(*)$$

as shown in the following diagram.

CRITICAL VALUES OF F-DISTRIBUTION



From Remark to Example 14·17, we have the following reciprocal relation between the upper and lower α-significant points of F-distribution :

$$F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)}$$

$$\Rightarrow \qquad F_\alpha(n_1, n_2) \times F_{1-\alpha}(n_2, n_1) = 1 \qquad \ldots(**)$$

The critical values of F for left tail test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 < \sigma_2^2$ are given by $F < F_{n_1-1,\, n_2-1}(1-\alpha)$, and for the two tailed test, $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$ are given by $F > F_{n_1-1,\, n_2-1}(\alpha/2)$ and $F < F_{n_1-1, n_2-1}(1 - \alpha/2)$ [For details, see § 16·7·5].

Example 14·20. *Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0·8 and 0·5 respectively. Assuming that the weight distributions are normal, test the hypothesis that the true variances are equal, against the alternative that they are not, at the 10% level.* [*Assume that* $P(F_{10,8} \geq 3{\cdot}35) = 0{\cdot}05$ *and* $P(F_{8,10} \geq 3{\cdot}07) = 0{\cdot}05$].

Solution. We want to test *Null Hypothesis*, $H_0 : \sigma_X^2 = \sigma_Y^2$.

against the *Alternative Hypothesis*, $H_1 : \sigma_X^2 \neq \sigma_Y^2$ (Two-tailed).

We are given :

$$n_1 = 11,\ n_2 = 9,\ s_X = 0{\cdot}8 \text{ and } s_Y = 0{\cdot}5.$$

Under the null hypothesis, $H_0 : \sigma_X = \sigma_Y$, the statistic

$$F = \frac{S_X^2}{S_Y^2}$$

follows F-distribution with $(n_1 - 1, n_2 - 1)$ d.f.

Now $\qquad n_1 s_X^2 = (n_1 - 1) S_X^2$

$$\therefore \qquad S_X^2 = \left(\frac{n_1}{n_1 - 1}\right) s_X^2 = \left(\frac{11}{10}\right) \times (0{\cdot}8)^2 = 0{\cdot}704$$

Similarly, $$S_Y^2 = \left(\frac{n_2}{n_2 - 1}\right) s_Y^2 = \left(\frac{9}{8}\right) \times (0{\cdot}5)^2 = 0{\cdot}28125$$

$$\therefore \qquad F = \frac{0{\cdot}704}{0{\cdot}28125} = 2{\cdot}5$$

The significant values of F for two tailed test at level of significance $\alpha = 0{\cdot}10$ are :

$$\left.\begin{aligned} F &> F_{10,8}(\alpha/2) = F_{10,8}(0{\cdot}05) \\ \text{and } F &< F_{10,8}(1 - \alpha/2) = F_{10,8}(0{\cdot}95) \end{aligned}\right\} \qquad \ldots(*)$$

We are given the tabulated (significant) values :

$$P[F_{10,8} \geq 3{\cdot}35] = 0{\cdot}05 \quad \Rightarrow \quad F_{10,8}(0{\cdot}05) = 3{\cdot}35 \qquad \ldots(**)$$

Also $$P[F_{8,10} \geq 3{\cdot}07] = 0{\cdot}05 \quad \Rightarrow \quad P\left[\frac{1}{F_{8,10}} \leq \frac{1}{3{\cdot}07}\right] = 0{\cdot}05$$

$$\Rightarrow \quad P[F_{10,8} \leq 0{\cdot}326] = 0{\cdot}05 \quad \Rightarrow \quad P[F_{10,8} \geq 0{\cdot}326] = 0{\cdot}95 \ldots(***)$$

Hence from (*), (**) and (***), the critical values for testing $H_0 : \sigma_X^2 = \sigma_Y^2$, against $H_1 : \sigma_X^2 \neq \sigma_Y^2$ at level of significance $\alpha = 0{\cdot}10$ are given by :

$$F > 3{\cdot}35 \quad \text{and} \quad F < 0{\cdot}326 = 0{\cdot}33$$

Since, the calculated value of F (=2·5) lies between 0·33 and 3·35, it is not significant and hence null hypothesis of equality of population variances may be accepted at level of significance $\alpha = 0{\cdot}10$.

Example 14·21. *In one sample of 8 observations, the sum of the squares of deviations of the sample values from the sample mean was 84·4 and in the other sample of 10 observations it was 102·6. Test whether this difference is significant at 5 per cent level, given that the 5 per cent point of F for $n_1 = 7$ and $n_2 = 9$ degrees of freedom is 3·29.* **[*Delhi Univ. B.Sc. (Maths Hons.), 1986*]**

Solution. Here $n_1 = 8,\ n_2 = 10$

and $$\Sigma(x - \bar{x})^2 = 84{\cdot}4, \quad \Sigma(y - \bar{y})^2 = 102{\cdot}6$$

$$\therefore \qquad S_X^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{84{\cdot}4}{7} = 12{\cdot}057$$

$$S_Y^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{102{\cdot}6}{9} = 11{\cdot}4$$

Under $H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$, *i.e.*, the estimates of σ^2 given by the samples are homogeneous, the test statistic is

$$F = \frac{S_X^2}{S_Y^2} = \frac{12{\cdot}057}{11{\cdot}4} = 1{\cdot}057$$

Tabulated $F_{0{\cdot}05}$ for (7, 9) d.f. is 3·29.

Since calculated $F < F_{0{\cdot}05}$, H_0 may be accepted at 5% level of significance.

Example 14·22. *Two random samples gave the following results :*

Sample	*Size*	*Sample mean*	*Sum of squares of deviations from the mean*
1	*10*	*15*	*90*
2	*12*	*14*	*108*

Test whether the samples come from the same normal population at 5% level of significance.

[*Given* : $F_{0{\cdot}05}$ (9, 11) = 2·90, $F_{0{\cdot}05}$(11,9) = 3·10 (approx.)

and $t_{0{\cdot}05}(20) = 2{\cdot}086$, $t_{0{\cdot}05}(22) = 2{\cdot}07$]

[Delhi Univ. MCA, 1987]

Solution. A normal population has two parameters, *viz.*, mean μ and variance σ^2. To test if two independent samples have been drawn from the same normal population we have to test (*i*) the equality of population means, and (*ii*) the equality of population variances.

Null Hypothesis : The two samples have been drawn from the same normal population, *i.e.*, $H_0 : \mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$.

Equality of means will be tested by applying t-test and equality of variances will be tested by applying F-test. Since t-test assumes $\sigma_1^2 = \sigma_2^2$, we shall first apply F-test and then t-test.

We are given $$\left.\begin{array}{l} n_1 = 10,\ n_2 = 12;\ \bar{x}_1 = 15,\ \bar{x}_2 = 14 \\ \Sigma(x_1 - \bar{x}_1)^2 = 90,\ \Sigma(x_2 - \bar{x}_2)^2 = 108 \end{array}\right\}$$

F-test

Here $$S_1^2 = \frac{1}{n_1 - 1}\Sigma(x_1 - \bar{x}_1)^2 = \frac{90}{9} = 10$$

$$S_2^2 = \frac{1}{n_2 - 1}\Sigma(x_2 - \bar{x}_2)^2 = \frac{108}{11} = 9{\cdot}82$$

Since $S_1^2 > S_2^2$, under $H_0 : \sigma_1^2 = \sigma_2^2$, the test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F\,(n_1 - 1, n_2 - 1) = F(9, 11)$$

Now $F = \dfrac{10}{9 \cdot 82} = 1 \cdot 018$

Tabulated $F_{0 \cdot 05}(9, 11) = 2 \cdot 90$

Since calculated F is less than tabulated F it is not significant. Hence null hypothesis of equality of population variances may be accepted.

Since $\sigma_1^2 = \sigma_2^2$, we can now apply t test for testing $H_0 : \mu_1 = \mu_2$.

t-test. Under $H_0' : \mu_1 = \mu_2$, against alternative hypothesis, $H_1' : \mu_1 \neq \mu_2$, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2} = t_{20}$$

where
$$S^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2]$$

$$= \frac{1}{20} [90 + 108] = 9 \cdot 9$$

$$\therefore \quad t = \frac{15 - 14}{\sqrt{9 \cdot 9 \left(\frac{1}{10} + \frac{1}{12} \right)}} = \frac{1}{\sqrt{9 \cdot 9 \times \frac{11}{60}}}$$

$$= \frac{1}{\sqrt{1 \cdot 815}} = 0 \cdot 742$$

Now $t_{0 \cdot 05}$ for 20 $d.f.$ $= 2 \cdot 086$

Since $|t| < t_{0 \cdot 05}$, it is not significant. Hence the hypothesis $H_0' : \mu_1 = \mu_2$ may be accepted. Since both the hypotheses, *i.e.*, $H_0' : \mu_1 = \mu_2$ and $H_0 : \sigma_1^2 = \sigma_2^2$ are accepted, we may regard that the given samples have been drawn from the same normal population.

EXERCISE 14(f)

1. (*a*) If χ_1^2 and χ_2^2 are independent chi-square variates with n_1 and n_2 d.f., obtain the probability density function of F-statistic defined by

$$F = \frac{(\chi_1^2 / n_1)}{(\chi_2^2 / n_2)}$$

Mention the types of hypotheses which are tested with the help of this statistic.

(*b*) Explain why the larger variance is placed in the numerator of the statistic F. Discuss the application of F-test in testing if two variances are homogeneous.

2. An investigator, newly appointed, was made to take ten independent measurements on the maximum internal diameter of a pot at specified equal intervals of time and the standard deviation of these ten observations was found

to be 0·0345 mm. After he had been some time on similar jobs, he was asked to repeat this experiment an equal number of times and the standard deviation of the new set of ten observations was found to be 0·0285 mm. Can it be concluded that the investigator has become more consistent (*i.e.* less variable) with practice ?

3. (*a*) Two independent samples of 8 and 7 items respectively had the following values of the variables.

Sample I	:	9	11	13	11	15	9	12	14
Sample II	:	10	12	10	14	9	8	10	

Do the estimates of population variance differ significantly ?

[*Delhi Univ. B.Sc., 1992*]

(*b*) Five measurements of the output of two units have given the following results (in kilograms of material per one hour of operation).

Unit A	:	14·1	10·1	14·7	13·7	14·0
Unit B	:	14·0	14·5	13·7	12·7	14·1

Assuming that both samples have been obtained from normal populations, test at 10% significance level if the two populations have the same variance, it being given that $F_{0.95}(4, 4) = 6.39$

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1991*]

(*c*) In one sample of 10 observations from a normal population, the sum of the squares of the deviations of the sample values from the sample mean is 102·4 and in another sample of 12 observations from another normal population, the sum of the squares of the deviations of the sample values from the sample mean is 120·5. Examine whether the two normal populations have the same variance.

4. (*a*) Two random samples of sizes 8 and 11, drawn from two normal populations, are characterised as follows ;

Population from which the sample is drawn	*Size of sample*	*Sum of observations*	*Sum of squares of observations*
I	8	9·6	61·52
II	11	16·5	73·26

You are to decide if the two populations can be taken to have the same variance. What test function would you use ? How is it distributed and what value it has in this sampling experiment ?

(*b*) The following are the values in thousands of an inch obtained by two engineers in 10 successive measurements with the same micrometer. Is one engineer significantly more consistent than the other ?

Engineer A : 503, 505, 497, 505, 495, 502, 499, 493, 510, 501
Engineer B : 502, 497, 492, 498, 499, 495, 497, 496, 498,

Ans. $H_0 : \sigma_1^2 = \sigma_2^2$ (both engineers are equally consistent). $F = 2.4$. Not significant.

(*c*) The nicotine content (in miligrams) of two samples of tobacco were found to be as follows :

Sample A	:	24	27	26	21	25	
Sample B	:	27	30	28	31	22	36

Can it be said that the two samples come from the same normal population ?

Ans. $H_0: \mu_1 = \mu_2$; $t = 1{\cdot}9$, Not significant.

$H_0' : \sigma_1^2 = \sigma_2^2$, $F = 4{\cdot}08 < 6{\cdot}26$ [$F_{0{\cdot}05}(5, 4)$]. Not significant.

Hence the two samples have come from the same normal population.

5. (*a*) Two random samples drawn from two normal populations are :

Sample I :	20,	16,	26,	27,	23,	22,	18,	24,	25,	19		
Sample II :	27,	33,	42,	35,	32,	34,	38,	28,	41,	43,	30,	37

Obtain estimates of the variances of the populations and test whether the populations have same variances.

[Given $F_{0{\cdot}05} = 3{\cdot}11$ for 11 and 9 degrees of freedom.]

(*b*) Test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$

given $n_1 = 25$, $\sum (x_i - \bar{x})^2 = 164 \times 24$,

$n_2 = 21$, $\sum (y_j - \bar{y})^2 = 190 \times 21$.

Make necessary assumptions, stating them.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1987*]

(*c*) The diameters of two random samples, each of size 10, of bullets produced by two machines have standard deviations $s_1 = 0{\cdot}01$ and $s_2 = 0{\cdot}015$. Assuming that the diameters have independent distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, test the hypothesis that, the two machines are equally good by testing :

$$H_0 : \sigma_1 = \sigma_2 \text{ against } H_1 : \sigma_1 \neq \sigma_2.$$

6. The following table shows the yield of corn in bushels per plot in 20 plots, half of which are treated with phosphate as fertiliser.

Treated :	5	0	8	3	6	1	0	3	3	1
Untreated :	1	4	1	2	3	2	5	0	2	0

Test whether the treatment by phosphate has

(*i*) changed the variability of the plot yields,

(*ii*) improved the average yield of corn.

7. (*a*) The following figures give the prices in rupees of a certain commodity in a sample of 15 shops selected at random from a city *A* and those in a sample of 13 shops from another city *B*.

City *A* :	7·41	7·77	7·44	7·40	7·38	7·93	7·58	
	8·28	7·23	7·52	7·82	7·71	7·84	7·63	7·68
City *B*	7·08	7·49	7·42	7·04	6·92	7·22	7·68	
	7·24	7·74	7·81	7·28	7·43	7·47		

Assuming that the distribution of prices in the two cities is normal, answer the following :

(*i*) Is it possible that the average price of city *B* is Rs. 7·20 ?

(*ii*) Is the observed variance in the first sample consistent with the hypothesis that the standard deviation of prices in city *A* is Rs. 0·30 ?

(*iii*) Is it reasonable to say that the variability of prices in the two cities is the same ?

(*iv*) Is it reasonable to say that the average prices are the same in two cities ?

14·5·6. Relation between t and F distributions. In F-distribution with (ν_1, ν_2) d.f. [c.f. 14·5 (*a*)], take $\nu_1 = 1$, $\nu_2 = \nu$ and $t^2 = F$, *i.e.*, $dF = 2t\, dt$. Thus the probability differential of F transforms to

$$dG(t) = \frac{(1/\nu)^{1/2}}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{(t^2)^{\frac{1}{2}-1}}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu+1)/2}}\, 2t\, dt, \quad 0 \le t^2 < \infty$$

$$= \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu+1)/2}}\, dt, \; -\infty < t < \infty$$

the factor 2 disappearing since the total probability in the range $(-\infty, \infty)$ is unity. This is the probability function of Student's t-distribution with ν d.f. Hence we have the following relation between t and F distributions.

'If a statistic t follows Student's t distribution with n d.f., then t^2 follows Snedecor's F-distribution with (1, n) d.f. Symbolically,

$$\left.\begin{array}{ll} \text{if} & t \sim t_{(n)} \\ \text{then} & t^2 \sim F_{(1,\, n)} \end{array}\right\} \qquad \ldots(14{\cdot}19)$$

Aliter Proof of (14·19). If $\xi \sim N(0, 1)$ and $X \sim \chi^2_{(n)}$ are independent *r.v.*'s then :

$$U = \xi^2 \sim \chi^2_{(1)} \qquad \text{[Square of a S.N.V.]}$$

and $$t = \frac{\xi}{\sqrt{X/n}} \sim t_{(n)}$$

$$\Rightarrow \quad t^2 = \frac{\xi^2}{(X/n)} = \frac{(\xi^2/1)}{(X/n)},$$

being the ratio of two independent chi-square variates divided by their respective degrees of freedom is $F(1, n)$ variate.

Hence $\quad t^2 \sim F(1, n)$

With the help of relation (14·19), all the uses of t-distribution can be regarded as the applications of F-distribution also, *e.g.*, for test for a single mean, instead of computing

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

we may compute

$$F = t^2 = \frac{n(\bar{x} - \mu)^2}{S^2}$$

and then apply F-test with $(1, n)$ d.f. and so on.

Similarly,we can write the test statistic F from § 14·2·9, § 14·2·10 and § 14·2·11 for testing the significance of an observed sample correlation coefficient, regression coefficient and partial correlation coefficient respectively.

Example 14·23. *Given* : $P[F(10, 12) > 2{\cdot}753] = 0{\cdot}05$
$= P[F(1, 12) > 4{\cdot}747]$

find $P[F(12, 10) > (2{\cdot}753)^{-1}]$, *and* $P[-\sqrt{4{\cdot}747} < t_{12} < \sqrt{4{\cdot}747}\,]$

Solution.

$$\begin{aligned} P[F(12, 10) > (2{\cdot}753)^{-1}] &= P\left[\frac{1}{F(12, 10)} < 2{\cdot}753\right] \\ &= P[F(10, 12) < 2{\cdot}753] \\ &= 1 - P[F(10, 12) > 2{\cdot}753] \\ &= 1 - 0{\cdot}05 = 0{\cdot}95 \\ P[-\sqrt{4{\cdot}747} < t_{12} < \sqrt{4{\cdot}747}\,] &= P(t^2_{12} < 4{\cdot}747) \\ &= P[F(1, 12) < 4{\cdot}747)] \\ &= 1 - P[F(1, 12) > 4{\cdot}747] \\ &= 1 - 0{\cdot}05 = 0{\cdot}95 \end{aligned}$$

14·5·7. Relation between F and χ^2. In $F(n_1, n_2)$ distribution if we let $n_2 \to \infty$, then $\chi^2 = n_1 F$ follows χ^2-distribution with n_1 d.f.

Proof. We have

$$p(F) = \frac{(n_1/n_2)^{n_1/2} F^{(n_1/2)-1}}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot \frac{\Gamma(n_1 + n_2)/2]}{\left[1 + \dfrac{n_1}{n_2} F\right]^{(n_1+n_2)/2}}, \quad 0 < F < \infty$$

In the limit as $n_2 \to \infty$, we have

$$\frac{\Gamma(n_1 + n_2)/2]}{n_2^{n_1/2}\,\Gamma(n_2/2)} \to \frac{(n_2/2)^{n_1/2}}{n_2^{n_1/2}} = \frac{1}{2^{n_1/2}}$$

$$\left[\because \frac{\Gamma(n+k)}{\Gamma(n)} \to n^k \text{ as } n \to \infty. (\textit{c.f.} \text{ Remark below.})\right]$$

Also
$$\lim_{n_2 \to \infty} \left[1 + \frac{n_1}{n_2} F\right]^{(n_1+n_2)/2} = \lim_{n_2 \to \infty} \left[\left(1 + \frac{n_1}{n_2} F\right)^{n_2}\right]^{1/2} \times \lim_{n_2 \to \infty} \left(1 + \frac{n_1}{n_2} F\right)^{n_1/2}$$

$$= \exp(n_1 F/2) = \exp(\chi^2/2) \qquad (\because n_1 F = \chi^2)$$

Hence in the limit, the p.d.f. of $\chi^2 = n_1 F$ becomes

$$\begin{aligned} dP(\chi^2) &= \frac{(n_1/2)^{n_1/2}\, e^{-\chi^2/2}}{\Gamma(n_1/2)} \cdot \left(\frac{\chi^2}{n_1}\right)^{(n_1/2)-1} d\left(\frac{\chi^2}{n_1}\right) \\ &= \frac{1}{2^{n_1/2}\,\Gamma(n_1/2)} \cdot e^{-\chi^2/2} (\chi^2)^{(n_1/2)-1}\, d\chi^2, \; 0 < \chi^2 < \infty \end{aligned}$$

which is the p.d.f. of chi-square distribution with n_1 d.f.

Remark. $\lim\limits_{n \to \infty} \dfrac{\Gamma(n+k)}{\Gamma(n)} = \lim\limits_{n \to \infty} \dfrac{(n+k-1)\,!}{(n-1)\,!}$, (for large n)

$$\simeq \lim \frac{\sqrt{2\pi}\, e^{-(n+k-1)}\, (n+k-1)^{n+k-(1/2)}}{\sqrt{2\pi}\, e^{-(n-1)}\, (n-1)^{n-(1/2)}}$$

(On using Stirling's approximation for n ! as $n \to \infty$.)

$$= e^{-k} \lim_{n \to \infty} \frac{n^{n+k-\frac{1}{2}} \left(1 + \dfrac{k-1}{n}\right)^{n+k-\frac{1}{2}}}{n^{n-\frac{1}{2}} \left(1 - \dfrac{1}{n}\right)^{n-\frac{1}{2}}}$$

$$= e^{-k} n^k \frac{\lim\limits_{n \to \infty} \left(1 + \dfrac{k-1}{n}\right)^n \lim\limits_{n \to \infty} \left(1 + \dfrac{k-1}{n}\right)^{k-\frac{1}{2}}}{\lim\limits_{n \to \infty} \left(1 - \dfrac{1}{n}\right)^n \lim\limits_{n \to \infty} \left(1 - \dfrac{1}{n}\right)^{-\frac{1}{2}}}$$

$$= e^{-k} n^k \left[\frac{e^{(k-1)}\,.\,1}{e^{-1}.\,1}\right] = n^k$$

$$\lim_{n \to \infty} \frac{\Gamma(n+k)}{\Gamma n} \doteq n^k$$

14·5·8. F-test for Testing the Significance of an Observed Multiple Correlation Coefficient. If R is the observed multiple correlation coefficient of a variate with k other variates in a random sample of size n from a $(k+1)$ variate population, then Prof. R.A. Fisher proved that under the null hypothesis (H_0) *that the multiple correlation coefficient in the population is zero, the statistic*

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} \qquad \ldots(14{\cdot}20)$$

conforms to F-distribution with $(k, n-k-1)$ d.f.

14·5·9. F-test for significance of an observed sample correlation Ratio η_{YX}. Under the null hypothesis that *population correlation ratio is zero,* the test statistic is

$$F = \frac{\eta^2}{1-\eta^2} \cdot \frac{N-h}{h-1} \sim F(h-1, N-h) \qquad \ldots(14{\cdot}21)$$

where N is the size of the sample (from a bi-variate normal population) arranged in h arrays.

14·5·10. F-test for Testing the Linearity of Regression. For a sample of size N arranged in h arrays, from a bi-variate normal population, the test statistic for testing the hypothesis of linearity of regression is

$$F = \frac{\eta^2 - r^2}{1-\eta^2} \, \frac{N-h}{h-2} \sim F(h-2, N-h) \qquad \ldots(14{\cdot}22)$$

14·5·11. F-test for Equality of Several Means. This test is carried out by the technique of Analysis of Variance, which plays a very important and fundamental role in Design of Experiments in Agricultural Statistics.

14·5. Non-Central F-distribution. The ratio of two independent χ^2 variates each divided by the corresponding d.f. has a non-central F-distribution if the numerator has a non-central χ^2-distribution and the denominator has a central χ^2-distribution. Thus, if X has a non-central χ^2-distribution with n_1 d.f. and non-centrality parameter λ, *i.e.*, if $X \sim \chi'^2_{n_1}$ and Y is an independent (central) χ^2-variate with n_2 d.f. *i.e.*, if $Y \sim \chi^2_{n_2}$, then the non-central F-statistic is defined as :

$$F' = \frac{X/n_1}{Y/n_2} = \frac{n_2 X}{n_1 Y}$$

p.d.f. of F′. Since X and Y are independent, their joint p.d.f. is given by

$$p(x, y) = p_1(x).\, p_2(y) = \left[\sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i\,!} \cdot \frac{e^{-x/2}\, x^{(n_1/2)+i-1}}{2^{(n_1+2i)/2}\,\Gamma[(n_1+2i)/2]}\right]$$

$$\times \frac{e^{-y/2}\, y^{(n_2/2)-1}}{2^{n_2/2}\,\Gamma(n_2/2)}\,;\quad 0 \le (x, y) < \infty.$$

Let us transform to the new set of r.v.'s F' and U defined by the transformation :

$$F' = \frac{n_2 x}{n_1 y} \quad \text{and} \quad u = y \quad \Rightarrow \quad y = u,\ x = \frac{n_1 u F'}{n_2}$$

$$J = \frac{\partial(x, y)}{\partial(F', u)} = \begin{vmatrix} \frac{n_1}{n_2} u & \frac{n_1}{n_2} F' \\ 0 & 1 \end{vmatrix} = \frac{n_1}{n_2} u$$

The joint p.d.f. of F' and U is given by

$$g(F', u) = \left\{\sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i\,!}\ \frac{\exp\left[-\frac{n_1 u F'}{2n_2}\right] \cdot \left(\frac{n u F'}{n_2}\right)^{(n_1/2)+i-1}}{2^{i+(n_1+n_2)/2}\,\Gamma(n_2/2)\,\Gamma[(n_1+2i)/2]}\right\}$$

$$\times e^{-u/2} u^{(n_2/2)-1} \cdot \left(\frac{n_1}{n_2}\right) u$$

$$= \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{\frac{e^{-\lambda}\lambda^i}{i\,!} \cdot \frac{\left(\frac{n_1}{n_2}F'\right)^{(n_1/2)+i-1}}{2^{i+(n_1+n_2)/2}\,\Gamma(n_2/2)\,\Gamma[(n_1+2i)/2]}\right.$$

$$\left. \times \exp\left[-\frac{u}{2}\left(1 + \frac{n_1}{n_2}F'\right)\right] \cdot u^{\frac{n_1+n_2}{2}+i-1}\right\}$$

$$0 \le F' < \infty,\ 0 < u < \infty$$

Integrating it w.r.t. u between the limits 0 to ∞ and using Gamma Integral, we obtain the marginal p.d.f. of F' as

$$g(F') = \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{ \frac{e^{-\lambda}\lambda^i}{i\,!} \cdot \frac{\left(\frac{n_1}{n_2}F'\right)^{(n_1/2)+i-1}}{2^{i+(n_1+n_2)/2}\,\Gamma(n_2/2)\,\Gamma[(n_1+2i)/2]} \times \frac{\Gamma\left(\frac{n_1+n_2}{2}+i\right)}{\left[\frac{1}{2}\left(1+\frac{n_1}{n_2}F'\right)\right]^{i+(n_1+n_2)/2}} \right\}$$

$$= \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{ \frac{e^{-\lambda}\lambda^i}{i\,!} \cdot \frac{\left(\frac{n_1}{n_2}F'\right)^{(n_1/2)+i-1}}{B\left(\frac{n_1}{2}+i, \frac{n_2}{2}\right)} \times \frac{1}{\left(1+\frac{n_1}{n_2}F'\right)^{i+(n_1+n_2)/2}} \right\} ; 0 \le F' < \infty \quad \ldots(14\cdot23)$$

Remarks. 1. For $\lambda = 0$, we get

$$g(F') = \frac{n_1}{n_2} \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{\left(\frac{n_1}{n_2}F'\right)^{(n_1/2)-1}}{\left(1+\frac{n_1}{n_2}F'\right)^{(n_1+n_2)/2}} ; 0 \le F' < \infty,$$

since for $\lambda = 0$, we get the contribution from the sum only when $i = 0$ and all other terms vanish. Thus for $\lambda = 0$, $g(F')$ reduces to the p.d.f. of central F-distribution with (n_1, n_2) d.f.

2. The hyper-geometric function of first kind is defined by

$$_1F_1(a, b, y) = \sum_{i=0}^{\infty} \frac{\Gamma(a+i)\,\Gamma b}{\Gamma a\,\Gamma(b+i)} \cdot \frac{y^i}{i\,!} \quad \ldots(14\cdot23b)$$

$$\therefore \quad {}_1F_1\left(\frac{n_1+n_2}{2}, \frac{n_1}{2}, \frac{\lambda n_1 F'}{n_2\left(1+\frac{n_1}{n_2}F'\right)}\right)$$

$$= \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{n_1+n_2}{2}+i\right)\Gamma\left(\frac{n_1}{2}\right)}{\Gamma\left(\frac{n_1+n_2}{2}\right)\Gamma\left(\frac{n_1}{2}+i\right)} \times \frac{(\lambda n_1 F')^i}{\left[n_2\left(1+\frac{n_1}{n_2}F'\right)\right]^i} \times \frac{1}{i\,!}$$

$$= \sum_{i=0}^{\infty} \frac{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)}{B\left(\frac{n_1}{2}+i, \frac{n_2}{2}\right)} \cdot \frac{\lambda^i}{i\,!} \frac{\left(\frac{n_1}{n_2}\right)^i F'^i}{\left(1+\frac{n_1}{n_2}F'\right)^i}$$

$$\therefore\ g(F') = \frac{e^{-\lambda} \cdot \left(\frac{n_1}{n_2}\right)^{n_1/2} (F')^{\frac{n_1}{2}-1}}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right) \cdot \left(1+\frac{n_1}{n_2}F'\right)^{\frac{n_1+n_2}{2}}} \times {}_1F_1\left(\frac{n_1+n_2}{2}, \frac{n_1}{2}, \frac{\lambda n_1 F'}{n_2\left(1+\frac{n_1}{n_2}F'\right)}\right)$$

3 It can be easily proved that the mean of F'_{n_1, n_2} is given by

$$E(F') = \int_0^{\infty} F' g(F')\, dF'$$

$$= \sum_{i=0}^{\infty} \left[\frac{e^{-\lambda}\lambda^i}{i\,!} \cdot \frac{n_2(n_1+2i)}{n_1(n_2-2)}\right];\ n_2 > 2. \qquad \text{...(14·23}c\text{)}$$

If $\lambda = 0$ (in which case we get contribution from the sum only when $i = 0$),

we get $$E(F') = \frac{n_2}{n_2-2}, \qquad \text{...(14·23}d\text{)}$$

which is the mean of central F-distribution with (n_1, n_2) d.f.

14·7. Fisher's z-distribution. In G.W. Snedecor's F-distribution with (ν_1, ν_2) d.f., if we put

$$F = \exp(2Z) \quad \Rightarrow \quad Z = \frac{1}{2}\log_e F \qquad \text{...(14·24)}$$

The distribution of Z becomes

$$g(z) = p(F) \cdot \left|\frac{dF}{dz}\right|$$

$$= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{(e^{2z})^{(\nu_1/2)-1}\, 2e^{2z}}{\left[1+\frac{\nu_1}{\nu_2}e^{2z}\right]^{(\nu_1+\nu_2)/2}}$$

$$= 2\frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{e^{\nu_1 z}}{\left[1+\frac{\nu_1}{\nu_2}e^{2z}\right]^{(\nu_1+\nu_2)/2}};\ -\infty < z < \infty \qquad \text{...(14·25)}$$

which is the probability function of Fisher's z-distribution with (ν_1, ν_2) d.f. The tables of significant values z_0 of z which will be exceeded in random sampling with probabilities 0·05 and 0·01, *i.e.*, $P(z > z_0) = 0·05$ and $P(z > z_0') = 0·01$

corresponding to various d.f. (ν_1, ν_2) were published by Fisher (*c.f.* Statistical Methods for Research Workers) in 1925. From these tables, Snedecor (1934-38) by using (14·24) deduced the tables of significant values of the variance ratio which he denoted by F in honour of Prof. R.A. Fisher.

Remark. With the help of relation (14·24), all the applications of F-distribution may be regarded as the applications of z-distribution also.

14·7·1. Moment Generating Function of z-distribution.

$$M_Z(t) = E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} g(z)\, dz$$

$$= \int_0^{\infty} F^{t/2} f(F)\, dF \qquad [\because e^{2z} = F]$$

Since μ_r' (about origin) for F-distribution is $\int_0^{\infty} F^r f(F)\, dF$, we can find m.g.f. of the z-distribution by putting $r = t/2$ in the expression for μ_r' for F-distribution.

Hence $\quad M_Z(t) = \left(\dfrac{\nu_2}{\nu_1}\right)^{t/2} \cdot \dfrac{\Gamma\{(\nu_1+t)/2\}\,\Gamma\{(\nu_2-t)/2\}}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}$ [*c.f.* Equation (14·15)]

$$\Rightarrow \quad K_Z(t) = \log M_Z(t)$$

$$= \frac{t}{2}[\log \nu_2 - \log \nu_1] + \log \Gamma\{(\nu_1+t)/2\}$$

$$+ \log \Gamma\{(\nu_2-t)/2\} - \log \Gamma(\nu_1/2) - \log \Gamma(\nu_2/2)$$

Using Stirling's approximation for $n!$, when n is large, *viz.*,

$$\lim_{n\to\infty} \Gamma(n+1) = \lim_{n\to\infty} n! \approx \sqrt{2\pi}\, e^{-n} n^{n+\frac{1}{2}}$$

$$\Rightarrow \quad \log \Gamma(n+1) = (n+\tfrac{1}{2}) \log n - n + \log \sqrt{2\pi}, \text{ we get}$$

$$\kappa_1 = \mu_1' = \frac{1}{2}\left(\frac{1}{\nu_2} - \frac{1}{\nu_1}\right)$$

$$\kappa_2 = \mu_2 = \frac{1}{2}\left(\frac{1}{\nu_2} + \frac{1}{\nu_1} + \frac{1}{\nu_1^2} + \frac{1}{\nu_2^2}\right)$$

$$\kappa_3 = \mu_3 = \frac{1}{2}\left[\left(\frac{1}{\nu_2^2} - \frac{1}{\nu_1^2}\right) + \left(\frac{1}{\nu_2^3} + \frac{1}{\nu_1^3}\right)\right]$$

$$\kappa_4 = \mu_4 - 3\mu_2^2 = \frac{1}{\nu_1^3} + \frac{1}{\nu_2^3} + 3\left(\frac{1}{\nu_1^4} + \frac{1}{\nu_2^4}\right),$$

whence β_1 and β_2 can be found.

Remark. z-distribution tends to normal distribution with mean $\frac{1}{2}\left(\frac{1}{\nu_2}-\frac{1}{\nu_1}\right)$ and variance $\frac{1}{2}\left(\frac{1}{\nu_1}+\frac{1}{\nu_2}\right)$, as ν_1 and ν_2 become large.

14·8. Fisher's z-transformation. To test the significance of an observed sample correlation coefficient from an uncorrelated bivariate normal population, t-test (*cf.* § 14·2·10) is used. But in random sample of size n, from a bivariate normal population in which $\rho \neq 0$, Prof. R.A. Fisher proved that the distribution of 'r' is by no means normal and in the neighbourhood of $\rho = \pm 1$, its probability curve is extremely skewed even for large n. If $\rho \neq 0$, Fisher suggested the following transformation

$$Z = \frac{1}{2}\log_e\frac{1+r}{1-r} = \tanh^{-1} r \qquad \ldots(14\cdot26)$$

and proved that even for small samples, the distribution of Z is approximately normal with mean

$$\xi = \frac{1}{2}\log_e\frac{1+\rho}{1-\rho} = \tanh^{-1}\rho \qquad \ldots[14\cdot26(a)]$$

and variance $1/(n-3)$ and for large values of n, say > 50, the approximation is fairly good.

z-transformation has the following applications in Statistics.

(1) To test if an observed value of 'r' differs significantly from a hypothetical value ρ of the population correlation coefficient.

H_0 : *There is no significant difference between r and ρ. In other words, the given sample has been drawn from a bivariate normal population with correlation coefficient ρ.*

If we take

$$Z = \tfrac{1}{2}\log_e\{(1+r)/(1-r)\} \text{ and } \xi = \tfrac{1}{2}\log_e\{(1+\rho)/(1-\rho)\},$$

then under H_0,

$$Z \sim N\left(\xi, \frac{1}{n-3}\right) \Rightarrow \frac{Z-\xi}{\sqrt{1/(n-3)}} \sim N(0, 1)$$

Thus if $(Z-\xi)\sqrt{(n-3)} > 1{\cdot}96$, H_0 is rejected at 5% level of significance and if it is greater than 2·58, H_0 is rejected at 1% level of significance, where Z and ξ are defined in (14·26) and (14·26a).

Remark. Z defined in equation (14·26) should not be confused with the Z used in Fisher's z-distribution (*c.f.* § 14·7).

Example 14·24. *A correlation coefficient of 0·72 is obtained from a sample of 29 pairs of observations.*

(i) Can the sample be regarded as drawn from a bivariate normal population in which true correlation coefficient is 0·8 ?

(ii) Obtain 95% confidence limits for ρ in the light of the information provided by the sample.

Solution. (*i*) H_0 : There is no significant difference between $r = 0{\cdot}72$; and

$\rho = 0{\cdot}80$, *i.e.*, the sample can be regarded as drawn from the bivariate normal population with $\rho = 0{\cdot}8$.

Here
$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1{\cdot}1513\log_{10}\left(\frac{1+r}{1-r}\right)$$
$$= 1{\cdot}1513\log_{10} 6{\cdot}14 = 0{\cdot}907$$
$$\xi = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right) = 1{\cdot}1513\log_{10}\frac{(1+0{\cdot}8)}{(1-0{\cdot}8)}$$
$$= 1{\cdot}1513 \times 0{\cdot}9541 = 1{\cdot}1$$
$$\text{S.E.}(Z) = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{26}} = 0{\cdot}196$$

Under H_0, the test statistic is
$$U = \frac{Z-\xi}{1/\sqrt{n-3}} \sim N(0, 1)$$

Now
$$U = \frac{(0{\cdot}907 - 1{\cdot}100)}{0{\cdot}196} = -0{\cdot}985$$

Since $|U| < 1{\cdot}96$, it is not significant at 5% level of significance and H_0 may be accepted. Hence the sample may be regarded as coming from a bivariate normal population with $\rho = 0{\cdot}8$.

(*ii*) 95% confidence limits for ρ on the basis of the information supplied by the sample, are given by
$$|U| \le 1{\cdot}96$$
$$|Z - \xi| \le 1{\cdot}96 \times \frac{1}{\sqrt{n-3}} = 1{\cdot}96 \times 0{\cdot}196$$
$$\Rightarrow \quad |0{\cdot}907 - \xi| \le 0{\cdot}384$$
$$\Rightarrow \quad 0{\cdot}907 - 0{\cdot}384 \le \xi \le 0{\cdot}907 + 0{\cdot}384$$
$$\Rightarrow \quad 0{\cdot}523 \le \xi \le 1{\cdot}291$$
$$\Rightarrow \quad 0{\cdot}523 \le \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right) \le 1{\cdot}291$$
$$\Rightarrow \quad 0{\cdot}523 \le 0{\cdot}1513\log_{10}\left(\frac{1+\rho}{1-\rho}\right) \le 1{\cdot}291$$
$$\Rightarrow \quad \frac{0{\cdot}523}{1{\cdot}1513} \le \log_{10}\left(\frac{1+\rho}{1-\rho}\right) \le \frac{1{\cdot}291}{1{\cdot}1513}$$
$$\Rightarrow \quad 0{\cdot}4543 \le \log_{10}\left(\frac{1+\rho}{1-\rho}\right) \le 1{\cdot}1213 \qquad \ldots(*)$$

Now
$$\log_{10}\left(\frac{1+\rho}{1-\rho}\right) = 0{\cdot}4543$$
$$\Rightarrow \quad \frac{1+\rho}{1-\rho} = \text{Antilog}(0{\cdot}4543) = 2{\cdot}846$$
$$\Rightarrow \quad \rho = \frac{2{\cdot}846 - 1}{2{\cdot}846 + 1} = \frac{1{\cdot}846}{3{\cdot}846} = 0{\cdot}4799$$

and
$$\log_{10}\left(\frac{1+\rho}{1-\rho}\right) = 1{\cdot}1213$$
$$\Rightarrow \quad \frac{1+\rho}{1-\rho} = \text{Antilog}(1{\cdot}1213) = 13{\cdot}22$$
$$\Rightarrow \quad \rho = \frac{13{\cdot}22 - 1}{13{\cdot}22 + 1} = \frac{12{\cdot}22}{14{\cdot}22} = 0{\cdot}86$$

Hence, substituting in (*) we get $0{\cdot}48 \leq \rho \leq 0{\cdot}86$

(2) **To test the significance of the difference between two independent sample correlation coefficients.** Let r_1 and r_2 be the sample correlation coefficients observed in two independent samples of sizes n_1 and n_2 respectively then

$$Z_1 = \frac{1}{2}\log_e\left(\frac{1+r_1}{1-r_1}\right) \text{ and } Z_2 = \frac{1}{2}\log_e\left(\frac{1+r_2}{1-r_2}\right)$$

Under the null hypothesis H_0 : *that sample correlation coefficients do not differ significantly, i.e., the samples are drawn from the same bivariate normal population or from different populations with same correlation coefficient ρ, (say), the statistic*

$$Z = \frac{(Z_1 - Z_2) - E(Z_1 - Z_2)}{\text{S.E.}(Z_1 - Z_2)} \sim N(0, 1)$$

Now $\quad E(Z_1 - Z_2) = E(Z_1) - E(Z_2) = \xi_1 - \xi_2 = 0$

$$\left[\because \xi_1 = \xi_2 = \frac{1}{2}\log_e\frac{1+\rho}{1-\rho} \text{ (under } H_0)\right]$$

and $\quad \text{S.E.}(Z_1 - Z_2) = \sqrt{V(Z_1) + V(Z_2)}$

[Covariance term vanishes since samples are independent]

$$= \sqrt{\left\{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right\}}$$

Under H_0, the test statistic is

$$Z = \frac{Z_1 - Z_2}{\sqrt{\left\{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right\}}} \sim N(0, 1)$$

By comparing this value with 1·96 or 2·58, H_0 may be accepted or rejected at 5% and 1% level of significance respectively.

(3) **To obtain pooled estimate of ρ.** Let $r_1, r_2, \ldots, r_k$ be observed correlation coefficients in k-independent samples of sizes $n_1, n_2, \ldots, n_k$ from a bivariate normal population. The problem is to combine these estimates of ρ to get a poled estimate for the parameter.

If we take $\quad Z_i = \frac{1}{2}\log_e\left(\frac{1+r_i}{1-r_i}\right);\ i = 1, 2, \ldots, k$

then Z_i ; $i = 1, 2, \ldots, k$ are independent normal variates with variances $\frac{1}{(n_i - 3)}$; $i = 1, 2, \ldots, k$ and common mean

$$\xi = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$$

The weighted mean (say $\bar{Z}$) of these Z_i's is given by

$$\bar{Z} = \sum_{i=1}^{k} w_i Z_i \Big/ \sum_{i=1}^{k} w_i\ ,$$

where w_i is the weight of Z_i.

Now $\bar{Z}$ is also an unbiased estimate of ξ, since

$$E(\bar{Z}) = \frac{1}{\Sigma w_i}\left[E \sum_{i=1}^{k} w_i Z_i\right] = \frac{1}{\Sigma w_i}\left[\sum_i w_i E(Z_i)\right] = \frac{1}{\Sigma w_i}\left[\sum_i w_i \xi\right] = \xi$$

and $V(\bar{Z}) = \dfrac{1}{(\Sigma w_i)^2} V[\Sigma w_i Z_i] = \dfrac{1}{(\Sigma w_i)^2}[\Sigma w_i^2 V(Z_i)]$

The weights w_i's, $(i = 1, 2, \ldots, n)$ are so chosen that $\bar{Z}$ has minimum variance.

In order that $V(\bar{Z})$ is minimum for variations in w_i, we should have

$$\frac{\partial}{\partial w_i} V(\bar{Z}) = 0 \ ; \quad i = 1, 2, \ldots, k$$

$$\Rightarrow \qquad \frac{(\Sigma w_i)^2 2 w_i V(Z_i) - [\sum_i w_i^2 V(Z_i)]\, 2(\sum_i w_i)}{(\Sigma w_i)^4} = 0$$

$$\Rightarrow \qquad w_i V(Z_i) = \frac{\Sigma w_i^2\, V(Z_i)}{\Sigma w_i}, \text{ a constant.}$$

$$\therefore \qquad w_i \propto \frac{1}{V(Z_i)} = (n_i - 3) \ ; \ i = 1, 2, \ldots, k \qquad \ldots(*)$$

Hence the minimum variance estimate of ξ is given by

$$\bar{Z} = \frac{\sum_{i=1}^{k} w_i Z_i}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k} (n_i - 3) Z_i}{\sum_{i=1}^{k} (n_i - 3)} \qquad \text{[On using (*)]}$$

and the best estimate of ρ is then given by

$$\bar{Z} = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} \quad \Rightarrow \quad \hat{\rho} = \tanh\left[\frac{\Sigma (n_i - 3) Z_i}{\Sigma (n_i - 3)}\right] \qquad (c.f. \ \S\ 13\cdot9\cdot1)$$

Remark. Minimum variance of $\bar{Z}$ is given by

$$[V(\bar{Z})]_{min} = \frac{\Sigma\left\{(n_i - 3)^2 \left(\dfrac{1}{n_i - 3}\right)\right\}}{\{\Sigma (n_i - 3)\}^2} = \frac{\Sigma (n_i - 3)}{\{\Sigma (n_i - 3)\}^2} = \frac{1}{\sum_{i=1}^{k} (n_i - 3)}$$

Statistical Inference-I (Theory of Estimation)

15·1. Introduction. The theory of estimation was founded by Prof. R.A. Fisher in a series of fundamental papers round about 1930.

Parameter Space. Let us consider a random variable X with p.d.f. $f(x, \theta)$. In most common applications, though not always, the functional form of the population distribution is assumed to be known except for the value of some unknown parameter(s) θ which may take any value on a set Θ. This is expressed by writing the p.d.f. in the form $f(x, \theta), \theta \in \Theta$. The set Θ, which is the set of all possible values of θ is called the *parameter space*. Such a situation gives rise not to one probability distribution but a family of probability distributions which we write as $\{f(x, \theta), \theta \in \Theta\}$. For example if $X \sim N(\mu, \sigma^2)$, then the parameter space

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty \,;\, 0 < \sigma < \infty\}$$

In particular, for $\sigma^2 = 1$, the family of probability distributions is given by

$$\{N(\mu, 1) \,;\, \mu \in \Theta\}, \text{ where } \Theta = \{\mu : -\infty < \mu < \infty\}$$

In the following discussion we shall consider a general family of distributions

$$\{f(x\,; \theta_1, \theta_2, \ldots, \theta_k) : \theta_i \in \Theta, i = 1, 2, \ldots, k\}.$$

Let us consider a random sample $x_1, x_2, \ldots, x_n$ of size n from a population, with probability function $f(x\,; \theta_1, \theta_2, \ldots, \theta_k)$, where $\theta_1, \theta_2, \ldots, \theta_k$ are the unknown population parameters. There will then always be an infinite number of functions of sample values, called statistics, which may be proposed as estimates of one or more of the parameters.

Evidently, the best estimate would be one that falls nearest to the true value of the parameter to be estimated. In other words, the statistic whose distribution concentrates as closely as possible near the true value of the parameter may be regarded the best estimate. Hence the basic problem of the estimation in the above case, can be formulated as follows :

'We wish to determine the functions of the sample observations :

$$T_1 = \hat{\theta}_1 (x_1, x_2, \ldots, x_n)\,,\, T_2 = \hat{\theta}_2 (x_1, x_2, \ldots, x_n), \ldots, T_k = \hat{\theta}_k\ (x_1, x_2, \ldots, x_n),$$

such that their distribution is concentrated as closely as possible near the true value of the parameter.

The estimating functions are then referred to as *estimators*.

15·2. Characteristics of Estimators. The following are some of the criteria that should be satisfied by a good estimator.

(*i*) *Consistency*

(ii) *Unbiasedness*

(iii) *Efficiency* and

(iv) *Sufficiency*

We shall now, briefly, explain these terms one by one.

15·3. Consistency. An estimator $T_n = T(x_1, x_2, \ldots, x_n)$, based on a random sample of size n, is said to be consistent estimator of $\gamma(\theta)$, $\theta \in \Theta$, the parameter space, if T_n converges to $\gamma(\theta)$ in probability.

i.e., if
$$T_n \xrightarrow{p} \gamma(\theta) \text{ as } n \to \infty \qquad \ldots(15\cdot1)$$

In other words, T_n is a consistent estimator of $\gamma(\theta)$ if for every $\varepsilon > 0$, $\eta > 0$, there exists a positive integer $n \geq m(\varepsilon, \eta)$ such that

$$P\left[|T_n - \gamma(\theta)| < \varepsilon\right] \to 1 \text{ as } n \to \infty \qquad \ldots(15\cdot2)$$

$$\Rightarrow \qquad P\left[|T_n - \gamma(\theta)| < \varepsilon\right] > 1 - \eta\ ;\ \forall\ n \geq m \qquad \ldots(15\cdot2a)$$

where m is some very large value of n.

Remark. If $X_1, X_2, \ldots, X_n$ is a random sample from a population with finite mean $EX_i = \mu < \infty$, then by Khinchine's weak law of large numbers (W.L.L.N), we have

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E(X_i) = \mu\ , \text{ as } n \to \infty.$$

Hence sample mean $(\bar{X}_n)$ is always a consistent estimator of the population mean (μ).

15·4. Unbiasedness. Obviously, consistency is a property concerning the behaviour of an estimator for indefinitely large values of the sample size n, *i.e.*, as $n \to \infty$. Nothing is regarded of its behaviour for finite n.

Moreover, if there exists a consistent estimator, say, T_n of $\gamma(\theta)$, then infinitely many such estimators can be constructed, *e.g.*,

$$T_n' = \left(\frac{n-a}{n-b}\right)T_n = \left[\frac{1-(a/n)}{1-(b/n)}\right]T_n \to T_n \xrightarrow{p} \gamma(\theta), \text{ as } n \to \infty$$

and hence, for different values of a and b, T_n' is also consistent for $\gamma(\theta)$.

Unbiasedness is a property associated with finite n. A statistic $T_n = T(x_1, x_2, \ldots, x_n)$, is said to be an unbiased estimator of $\gamma(\theta)$ if

$$E(T_n) = \gamma(\theta), \text{ for all } \theta \in \Theta \qquad \ldots(15\cdot3)$$

We have seen (c.f. § 12·12) that in sampling from a population with mean μ and variance σ^2,

$$E(\bar{x}) = \mu \text{ and } E(s^2) \neq \sigma^2 \text{ but } E(S^2) = \sigma^2.$$

Hence there is a reason to prefer

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ to the sample variance } s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Remark. If $E(T_n) > \theta$, T_n is said to be positively biased and if $E(T_n) < \theta$, it is said to be negatively biased, the amount of bias $b(\theta)$ being given by

$$b(\theta) = E(T_n) - \gamma(\theta),\ \theta \in \Theta \qquad \ldots(15\cdot3a)$$

15·4·1. Invariance Property of Consistent Estimators.

Theorem 15·1. *If T_n is a consistent estimator of $\gamma(\theta)$ and $\psi(\gamma(\theta))$ is a continuous function of $\gamma(\theta)$, then $\psi(T_n)$ is a consistent estimator of $\psi(\gamma(\theta))$.*

Proof. Since T_n is a consistent estimator of $\gamma(\theta)$, $T_n \xrightarrow{p} \gamma(\theta)$ as $n \to \infty$ *i.e.*, for every $\varepsilon > 0, \eta > 0, \exists$ a positive integer $n \geq m\ (\varepsilon, \eta)$ such that

$$P\left[|T_n - \gamma(\theta)| < \varepsilon\right] > 1 - \eta\ ,\ \forall\ n \geq m \qquad \ldots(*)$$

Since $\psi(\cdot)$ is a continuous function, for every $\varepsilon > 0$, however small, $\exists$ a positive number ε_1 such that $|\psi(T_n) - \psi(\gamma(\theta))| < \varepsilon_1$, whenever $|T_n - \gamma(\theta)| < \varepsilon$

i.e., $$|T_n - \gamma(\theta)| < \varepsilon\ \Rightarrow\ |\psi(T_n) - \psi(\gamma(\theta))| < \varepsilon_1 \qquad \ldots(**)$$

For two events A and B,

if $A \Rightarrow B$, then $A \subseteq B \ \Rightarrow\ P(A) \leq P(B) \ \Rightarrow\ P(B) \geq P(A) \qquad \ldots(***)$

From (**) and (***), we get

$$P\left[|\psi(T_n) - \psi(\gamma(\theta))| < \varepsilon_1\right] \geq P\left[|T_n - \gamma(\theta)| < \varepsilon\right]$$

$$\Rightarrow \qquad P\left[|\psi(T_n) - \psi(\gamma(\theta)| < \varepsilon_1\right] \geq 1 - \eta\ ;\ \forall\ n \geq m \qquad \text{[Using (*)]}$$

$$\Rightarrow \qquad \psi(T_n) \xrightarrow{p} \psi(\gamma(\theta)),\ \text{as } n \to \infty$$

$$\Rightarrow \qquad \psi(T_n) \text{ is a consistent estimator of } \gamma(\theta).$$

15·4·2. Sufficient Conditions for Consistency.

Theorem 15·2. Let $\{T_n\}$ *be a sequence of estimators such that for all $\theta \in \Theta$,*

(i) $E_\theta(T_n) \to \gamma(\theta),\ n \to \infty$

and *(ii)* $Var_\theta(T_n) \to 0$, *as* $n \to \infty$.

Then T_n is a consistent estimator of $\gamma(\theta)$.

Proof. We have to prove that T_n is a consistent estimator of $\gamma(\theta)$

i.e., $$T_n \xrightarrow{p} \gamma(\theta),\ \text{as } n \to \infty$$

i.e., $$P\left[|T_n - \gamma(\theta)| < \varepsilon\right] > 1 - \eta\ ;\ \forall\ n \geq m\ (\varepsilon, \eta) \qquad \ldots(15\cdot4)$$

where ε and η are arbitrarily small positive numbers and m is some large value of n:

Applying Chebychev's inequality to the statistic T_n, we get

$$P\left[|T_n - E_\theta(T_n)| \leq \delta\right] \geq 1 - \frac{\text{Var}_\theta(T_n)}{\delta^2} \qquad \ldots(15\cdot5)$$

We have

$$|T_n - \gamma(\theta)| = |T_n - E(T_n) + E(T_n) - \gamma(\theta)|$$

$$\leq |T_n - E_\theta(T_n)| + |E_\theta(T_n) - \gamma(\theta)| \qquad ...(15 \cdot 6)$$

Now

$$|T_n - E_\theta(T_n)| \leq \delta \Rightarrow |T_n - \gamma(\theta)| \leq \delta + |E_\theta(T_n) - \gamma(\theta)| \qquad ...(15 \cdot 7)$$

Hence, on using (***) of Theorem 15·1, we get

$$P\left[|T_n - \gamma(\theta)| \leq \delta + |E_\theta(T_n) - \gamma(\theta)|\right] \geq P\left[|T_n - E_\theta(T_n)| \leq \delta\right]$$

$$\geq 1 - \frac{\text{Var}_\theta(T_n)}{\delta^2} \quad \text{[From (15·5)]} \qquad ...(15 \cdot 8)$$

We are given :

$E_\theta(T_n) \to \gamma(\theta) \ \forall \ \theta \in \Theta$ as $n \to \infty$.

Hence, for every $\delta_1 > 0$, $\exists$ a positive integer $n \geq n_0(\delta_1)$ such that

$$|E_\theta(T_n) - \gamma(\theta)| \leq \delta_1, \ \forall \ n \geq n_0(\delta_1) \qquad ...(15 \cdot 9)$$

Also $\text{Var}_\theta(T_n) \to 0$ as $n \to \infty$, (Given).

$$\therefore \qquad \frac{\text{Var}_\theta(T_n)}{\delta^2} \leq \eta, \ \forall \ n \geq n_0'(\eta) \qquad ...(15 \cdot 10)$$

where η is arbitrarily small positive number.

Substituting from (15·9) and (15·10) in (15·8), we get

$$P\left[|T_n - \gamma(\theta)| \leq \delta + \delta_1\right] \geq 1 - \eta \ ; n \geq m(\delta_1, \eta)$$

$$\Rightarrow \qquad P\left[|T_n - \gamma(\theta)| \leq \varepsilon\right] \geq 1 - \eta \ ; n \geq m$$

where $m = \max(n_0, n_0')$ and $\varepsilon = \delta + \delta_1 > 0$.

$$\Rightarrow \qquad T_n \xrightarrow{p} \gamma(\theta), \text{ as } n \to \infty \qquad \text{[Using (15·4)]}$$

$\Rightarrow \qquad T_n$ is a consistent estimator of $\gamma(\theta)$.

Example 15·1. *$x_1, x_2, \ldots x_n$ is a random sample from a normal population $N(\mu, 1)$. Show that $t = \frac{1}{n} \sum_{i=1}^{n} x_i^2$, is an unbiased estimator of $\mu^2 + 1$.*

Solution. (*a*) We are given

$$E(x_i) = \mu, \ V(x_i) = 1 \ \forall \ i = 1, 2, \ldots, n$$

Now $\qquad E(x_i^2) = V(x_i) + \{E(x_i)\}^2 = 1 + \mu^2$

$$E(t) = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i^2\right] = \frac{1}{n} \sum_{i=1}^{n} E(x_i^2) = \frac{1}{n} \sum_{i=1}^{n} (1 + \mu^2) = 1 + \mu^2$$

Hence t is an unbiased estimator of $1 + \mu^2$.

Example 15·2. *If T is an unbiased estimator for θ, show that T^2 is a biased estimator for θ^2.*

Solution. Since T is an unbiased estimator for θ, we have

$$E(T) = \theta$$

Also $\qquad \text{Var}(T) = E(T^2) - [E(T)]^2 = E(T^2) - \theta^2$

$$\Rightarrow \qquad E(T^2) = \theta^2 + \text{Var}(T), \quad (\text{Var}\, T > 0).$$

Since $E(T^2) \neq \theta^2$, T^2 is a biased estimator for θ^2.

Example 15·3. *Show that* $\dfrac{[\Sigma x_i (\Sigma x_i - 1)]}{n(n-1)}$ *is an unbiased estimate of* θ^2, *for the sample* $x_1, x_2, \ldots, x_n$ *drawn on X which takes the values 1 or 0 with respective probabilities* θ *and* $(1-\theta)$.

Solution. Since $x_1, x_2, \ldots, x_n$ is a random sample from Bernoulli population with parameter θ,

$$T = \sum_{i=1}^{n} x_i \sim B(n, \theta)$$

$$\Rightarrow \qquad E(T) = n\theta \quad \text{and} \quad \text{Var}(T) = n\,\theta\,(1-\theta)$$

$$E\left[\frac{\Sigma x_i (\Sigma x_i - 1)}{n(n-1)}\right] = E\left[\frac{T(T-1)}{n(n-1)}\right]$$

$$= \frac{1}{n(n-1)}[E(T^2) - E(T)]$$

$$= \frac{1}{n(n-1)}\left[\text{Var}(T) + \{E(T)\}^2 - E(T)\right]$$

$$= \frac{1}{n(n-1)}\left[n\,\theta\,(1-\theta) + n^2\theta^2 - n\,\theta\right]$$

$$= \frac{n\,\theta^2\,(n-1)}{n(n-1)} = \theta^2$$

$\Rightarrow \quad [\Sigma x_i (\Sigma x_i - 1)] \,/\, [n(n-1)]$ is an unbiased estimator of θ^2.

Example 15·4. *Let X be distributed in the Poisson form with parameter* θ. *Show that the only unbiased estimator of* $\exp[-(k+1)\theta]$, $k > 0$, *is* $T(X) = (-k)^X$ *so that*

$$T(x) > 0 \text{ if } x \text{ is even}$$

and

$$T(x) < 0 \text{ if } x \text{ is odd.}$$

[***Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1988***]

Solution. $E\{T(X)\} = E\left[(-k)^X\right], k > 0 = \displaystyle\sum_{x=0}^{\infty} (-k)^x \left\{\frac{e^{-\theta}\,\theta^x}{x\,!}\right\}$

$$= e^{-\theta} \sum_{x=0}^{\infty} \left[\frac{(-k\theta)^x}{x\,!}\right] = e^{-\theta} \cdot e^{-k\theta} = e^{-(1+k)\theta}$$

$\Rightarrow \quad T(X) = (-k)^X$ is an unbiased estimator for $\exp\left[-(1+k)\,\theta\right]$, $k > 0$.

Example 15·5. (*a*) *Prove that in sampling from a* $N(\mu, \sigma^2)$ *population, the sample mean is a consistent estimator of* μ.

(*b*) *Prove that for Cauchy's distribution not sample mean but sample median is a consistent estimator of the population mean.*

Solution. In sampling from a $N(\mu, \sigma^2)$ population, the sample mean $\bar{x}$ is also normally distributed as $N(\mu, \sigma^2/n)$.

$$\Rightarrow \qquad E(\bar{x}) = \mu \quad \text{and} \quad V(\bar{x}) = \sigma^2/n$$

Thus as $n \to \infty$,

$$E(\bar{x}) = \mu \quad \text{and} \quad V(\bar{x}) = 0$$

Hence by Theorem 15·2, $\bar{x}$ is a consistent estimator for μ.

(*b*) The Cauchy's population is given by the probability function

$$dF(x) = \frac{1}{\pi} \cdot \frac{dx}{1 + (x - \mu)^2}, \; -\infty < x < \infty$$

The mean of the distribution, if we conventionally agree to assume that it exists, is at $x = \mu$.

If $\bar{x}$, the sample mean is taken as an estimator of μ, then the sampling distribution of $\bar{x}$ is given by

$$dF(\bar{x}) = \frac{1}{\pi} \cdot \frac{d\bar{x}}{1 + (\bar{x} - \mu)^2} \; ; -\infty < \bar{x} < \infty \qquad \ldots(i)$$

because in Cauchy's distribution, the distribution of $\bar{x}$ is same as the distribution of x.

Since in this case, the distribution of $\bar{x}$ is same as the distribution of any single sample observation, it does not increase in accuracy with increasing n. Hence we have

$$E(\bar{x}) = \mu \text{ but } V(\bar{x}) = V(x) \neq 0, \text{ as } n \to \infty$$

Hence by Theorem 15·2, $\bar{x}$ is not a consistent estimator of μ in this case.

Consideration of symmetry of (*i*) is enough to show that the sample median *Md* is an unbiased estimate of the population mean, which of course is same as the population median.

$$\therefore \qquad E(Md) = \mu \qquad \ldots(ii)$$

For large n, the sampling distribution of median is asymptotically normal and is given by

$$dF \propto \exp\left[-2n f_1^2 (x - \mu)^2\right] dx$$

where f_1 is the median ordinate of the parent population.

$$i.e., \qquad dF \propto \exp\left\{-\frac{(x - \mu)^2}{1/(2n f_1^2)}\right\} \qquad \ldots(iii)$$

But

$$\begin{aligned} f_1 &= \text{Median ordinate of } (i) \\ &= \text{Modal ordinate of } (i) \qquad \text{[Because of symmetry]} \\ &= [f(x)]_{x=\mu} = \frac{1}{\pi} \end{aligned}$$

Hence, from (*iii*), the variance of the sampling distribution of median is :

$$V(Md) = \frac{1}{4n f_1^2} = \frac{1}{4n(1/\pi)^2} = \frac{\pi^2}{4n} \to 0 \text{ as } n \to \infty \qquad \ldots(iv)$$

Hence from (*ii*) and (*iv*), using Theorem 15·2, we conclude that for Cauchy's distribution, median is a consistent estimator for μ.

Example 15·6. *If $X_1, X_2, \ldots, X_n$ are random observations on a Bernoulli variate X taking the value 1 with probability p and the value 0 with probability (1 – p), show that :*

$$\frac{\Sigma x_i}{n}\left(1 - \frac{\Sigma x_i}{n}\right) \text{ is a consistent estimator of } p(1-p).$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

Solution. Since $X_1, X_2, \ldots, X_n$ are *i.i.d* Bernoulli variates with parameter 'p',

$$T = \sum_{i=1}^{n} x_i \sim B(n, p)$$

$$\Rightarrow \quad E(T) = np \quad \text{and} \quad \text{Var}(T) = npq$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{T}{n}$$

$$\therefore \quad E(\bar{X}) = \frac{1}{n} E(T) = \frac{1}{n} \cdot np = p$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{T}{n}\right) = \frac{1}{n^2} \cdot \text{Var}(T) = \frac{pq}{n} \to 0 \text{ as } n \to \infty.$$

Since $E(\bar{X}) \to p$ and Var $(\bar{X}) \to 0$, as $n \to \infty$; $\bar{X}$ is a consistent estimator of p.

Also $\dfrac{\Sigma x_i}{n}\left(1 - \dfrac{\Sigma x_i}{n}\right) = \bar{X}\ (1 - \bar{X})$, being a polynomial in $\bar{X}$, is a continuous function of $\bar{X}$.

Since $\bar{X}$ is consistent estimator of p, by the invariance property of consistent estimators (Theorem 15·1), $\bar{X}\ (1 - \bar{X})$ is a consistent estimator of $p\ (1 - p)$.

15·5. Efficient Estimators. *Efficiency.* Even if we confine ourselves to unbiased estimates, there will, in general, exist more than one consistent estimator of a parameter. For example, in sampling from a normal population $N(\mu, \sigma^2)$, when σ^2 is known, sample mean $\bar{x}$ is an unbiased and consistent estimator of μ [*c.f.* Example 15·5*a*].

From symmetry it follows immediately that sample median (*Md*) is an unbiased estimate of μ, which is the same as the population median. Also for large n,

$$V(Md) = \frac{1}{4n f_1^2} \qquad [c.f. \text{ Example 15·5}(b)]$$

Here $\quad f_1$ = Median ordinate of the parent distribution.

= Modal ordinate of the parent distribution.

$$= \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x-\mu)^2/2\sigma^2\}\right]_{x=\mu} = \frac{1}{\sigma\sqrt{2\pi}}$$

$$\therefore \quad V(Md) = \frac{1}{4n} \cdot 2\pi\sigma^2 = \frac{\pi\sigma^2}{2n}$$

Since and $\left.\begin{array}{l} E(Md) = \mu \\ V(Md) \to 0 \end{array}\right\}$, as $n \to \infty$

median is also an unbiased and consistent estimator of μ.

Thus, there is a necessity of some further criterion which will enable us to choose between the estimators with the common property of consistency. Such a criterion which is based on the variances of the sampling distribution of estimators is usually known as *efficiency*.

If, of the two consistent estimators T_1, T_2 of a certain parameter θ, we have

$$V(T_1) < V(T_2), \text{ for all } n \qquad \ldots(15\cdot11)$$

then T_1 is more efficient than T_2 for all samples sizes.

We have seen above :

For all n, $\quad V(\bar{x}) = \dfrac{\sigma^2}{n}$

and for large n, $\quad V(Md) = \dfrac{\pi\sigma^2}{2n} = 1\cdot57\,\dfrac{\sigma^2}{n}$

Since $V(\bar{x}) < V(Md)$, we conclude that for normal distribution, sample mean is more efficient estimator for μ than the sample median, for large samples at least.

15·5·1. Most Efficient Estimator. *If in a class of consistent estimators for a parameter, there exists one whose sampling variance is less than that of any such estimator, it is called the most efficient estimator. Whenever such an estimator exists, it provides a criterion for measurement of efficiency of the other estimators.*

Efficiency *(Def.) If T_1 is the most efficient estimator with variance V_1 and T_2 is any other estimator with variance V_2, then the efficiency E of T_2 is defined as :*

$$E = \frac{V_1}{V_2} \qquad \ldots(15\cdot12)$$

Obviously, E cannot exceed unity.

If $T, T_1, T_2, \ldots, T_n$ are all estimators of $\gamma(\theta)$ and Var(T) is minimum, then the efficiency E_i of T_i, $(i = 1, 2, \ldots, n)$ is defined as :

$$E_i = \frac{\text{Var}\, T}{\text{Var}\, T_i}\,;\ i = 1, 2, \ldots, n \qquad \ldots(15\cdot12a)$$

Obviously $E_i \le 1, i = 1, 2, \dots n$.

For example, in the normal samples, since sample mean $\bar{x}$ is the most efficient estimator of μ [c.f. Remark to Example 15·31], the efficiency E of Md for such samples, (for large n), is :

$$E = \frac{V(\bar{x})}{V(Md)} = \frac{\sigma^2/n}{\pi\sigma^2/(2n)} = \frac{2}{\pi} = 0{\cdot}637$$

Example 15·7. *A random sample $(X_1, X_2, X_3, X_4, X_5)$ of size 5 is drawn from a normal population with unknown mean μ. Consider the following estimators to estimate μ :*

(i) $t_1 = \dfrac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$

(ii) $t_2 = \dfrac{X_1 + X_2}{2} + X_3$, (iii) $t_3 = \dfrac{2X_1 + X_2 + \lambda X_3}{3}$

where λ is such that t_3 is an unbiased estimator of μ.

Find λ. Are t_1 and t_2 unbiased ? State giving reasons, the estimator which is best among t_1, t_2 and t_3.

Solution. We are given

$E(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2$, (say) ; $\text{Cov}(X_i, X_j) = 0$, $(i \ne j = 1, 2, \dots, n)$ ...(*)

(i) $\quad E(t_1) = \dfrac{1}{5} \sum\limits_{i=1}^{5} E(X_i) = \dfrac{1}{5} \sum\limits_{i=1}^{5} \mu = \dfrac{1}{5} \cdot 5\mu = \mu$

$\Rightarrow$ t_1 is an unbiased estimator of μ.

(ii) $\quad E(t_2) = \frac{1}{2} E(X_1 + X_2) + E(X_3)$

$= \frac{1}{2}(\mu + \mu) + \mu$ [Using (*)]

$= 2\mu$

$\Rightarrow$ t_2 is *not* an unbiased estimator of μ.

(iii) $\quad E(t_3) = \mu$

$\Rightarrow \quad \frac{1}{3} E(2X_1 + X_2 + \lambda X_3) = \mu$

$\Rightarrow \quad 2E(X_1) + E(X_2) + \lambda E(X_3) = 3\mu$

$\Rightarrow \quad 2\mu + \mu + \lambda\mu = 3\mu$

$\Rightarrow \quad \lambda\mu = 0 \quad \Rightarrow \quad \lambda = 0$

Using (*), we get

$$V(t_1) = \frac{1}{25}[V(X_1) + V(X_2) + V(X_3) + V(X_4) + V(X_5)] = \frac{1}{5}\sigma^2$$

$$V(t_2) = \frac{1}{4}[V(X_1) + V(X_2)] + V(X_3) = \frac{1}{2}\sigma^2 + \sigma^2 = \frac{3}{2}\sigma^2$$

$$V(t_3) = \frac{1}{9}[4V(X_1) + V(X_2)] = \frac{1}{9}(4\sigma^2 + \sigma^2) = \frac{5}{9}\sigma^2 \qquad (\because \lambda = 0)$$

Since $V(t_1)$ is the least, t_1 is the best estimator (in the sense of least variance) of μ.

Example 15·8. *X_1, X_2, and X_3 is a random sample of size 3 from a population with mean value μ and variance σ^2, T_1, T_2, T_3 are the estimators used to estimate mean value μ, where*

$$T_1 = X_1 + X_2 - X_3, \quad T_2 = 2X_1 + 3X_3 - 4X_2, \quad \text{and } T_3 = (\lambda X_1 + X_2 + X_3)/3$$

(i) Are T_1 and T_2 unbiased estimators ?

(ii) Find the value of λ such that T_3 is unbiased estimator for μ.

(iii) With this value of λ is T_3 a consistent estimator ?

(iv) Which is the best estimator ?

Solution. Since X_1, X_2, X_3 is a random sample from a population with mean μ and variance σ^2,

$$E(X_i) = \mu, \ \text{Var}(X_i) = \sigma^2 \text{ and } \text{Cov}(X_i, X_j) = 0, \ (i \neq j = 1, 2, \ldots, n) \quad \ldots(*)$$

(*i*) $E(T_1) = E(X_1) + E(X_2) - E(X_3) = \mu + \mu - \mu = \mu$

$\Rightarrow$ T_1 is an unbiased estimator of μ

$E(T_2) = 2E(X_1) + 3E(X_3) - 4E(X_2) = 2\mu + 3\mu - 4\mu = \mu$

$\Rightarrow$ T_2 is an unbiased estimator of μ.

(*ii*) We are given : $E(T_3) = \mu$

$$\Rightarrow \quad \frac{1}{3}[\lambda E(X_1) + E(X_2) + E(X_3) = \mu$$

$$\Rightarrow \quad \frac{1}{3}(\lambda\mu + \mu + \mu) = \mu \ \Rightarrow \ \lambda\mu + 2\mu = 3\mu \ \Rightarrow \ \lambda = 1.$$

(*iii*) With $\lambda = 1$, $T_3 = \frac{1}{3}(X_1 + X_2 + X_3) = \overline{X}$

Since sample mean is a consistent estimator of population mean μ, by Weak Law of Large Numbers, T_3 is a consistent estimator of μ.

(*iv*) We have [on using (*)] :

$$\text{Var}(T_1) = \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) = 3\sigma^2$$

$$\text{Var}(T_2) = 4\,\text{Var}(X_1) + 9\,\text{Var}(X_3) + 16\,\text{Var}(X_2) = 29\,\sigma^2$$

$$\text{Var}(T_3) = \frac{1}{9}[\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)] = \frac{1}{3}\sigma^2 \qquad (\because \lambda = 1)$$

Since $\text{Var}(T_3)$ is minimum, T_3 is the best estimator in the sense of minimum variance.

15·5·2. Minimum Variance Unbiased (M.V.U.) Estimators. *If a statistic $T = T(x_1, x_2, \ldots, x_n)$, based on sample of size n is such that :*

(i) T is unbiased for $\gamma(\theta)$, for all $\theta \in \Theta$ and

(ii) It has the smallest variance among the class of all unbiased estimators of $\gamma(\theta)$,

then T is called the minimum variance unbiased estimator (MVUE) of $\gamma(\theta)$.

More precisely, T is MVUE of $\gamma(\theta)$ if

$$E_\theta(T) = \gamma(\theta) \text{ for all } \theta \in \Theta \quad \ldots(15\cdot13)$$

and $$\text{Var}_\theta(T) \le \text{Var}_\theta(T') \text{ for all } \theta \in \Theta \quad \ldots(15\cdot14)$$

where T' is any other unbiased estimator of $\gamma(\theta)$.

We give below some important Theorems concerning MVU estimators.

Theorem 15·3. *An M.V.U. is unique in the sense that if T_1 and T_2 are M.V.U. estimators for $\gamma(\theta)$, then $T_1 = T_2$, almost surely.*

Proof. We are given that

$$\left.\begin{aligned} E_\theta(T_1) &= E_\theta(T_2) = \gamma(\theta), \text{ for all } \theta \in \Theta \\ \text{and} \quad \text{Var}_\theta(T_1) &= \text{Var}_\theta(T_2) \text{ for all } \theta \in \Theta \end{aligned}\right\} \quad \ldots(15\cdot15)$$

Consider a new estimator

$$T = \frac{1}{2}(T_1 + T_2)$$

which is also unbiased since

$$E(T) = \frac{1}{2}[E(T_1) + E(T_2)] = \theta$$

$$\begin{aligned} \text{Var}(T) &= \text{Var}\left[\frac{1}{2}(T_1 + T_2)\right] = \frac{1}{4}\text{Var}(T_1 + T_2) \quad [\because \text{Var}(CX) = C^2\,\text{Var}(X)] \\ &= \frac{1}{4}[\text{Var}(T_1) + \text{Var}(T_2) + 2\,\text{Cov}(T_1, T_2)] \\ &= \frac{1}{4}[\text{Var}(T_1) + \text{Var}(T_2) + 2\rho\sqrt{\text{Var}(T_1)\,\text{Var}(T_2)}\,] \\ &= \frac{1}{2}\,\text{Var}(T_1)(1 + \rho), \qquad \ldots[\text{From } (15\cdot15)] \end{aligned}$$

where ρ is Karl Pearson's co-efficient of correlation between T_1 and T_2.

Since T_1 is the *MUV* estimator,

$$\begin{aligned} \text{Var}(T) &\ge \text{Var}(T_1) \\ \Rightarrow \quad \frac{1}{2}\text{Var}(T_1)[1 + \rho] &\ge \text{Var}(T_1) \\ \Rightarrow \quad \frac{1}{2}(1 + \rho) &\ge 1, \; i.e., \; \rho \ge 1 \end{aligned}$$

Since $|\rho| \le 1$, we must have $\rho = 1$, *i.e.*, T_1 and T_2 must have a linear relation of the form :

$$T_1 = \alpha + \beta T_2, \quad \ldots(15\cdot16)$$

where α and β are constants independent of $x_1, x_2, \ldots, x_n$ but may depend on θ, *i.e.*, we may have $\alpha = \alpha(\theta)$ and $\beta = \beta(\theta)$.

Taking expectation of both sides in (15·16) and using (15·15), we get

$$\theta = \alpha + \beta\theta \quad \ldots(15\cdot17)$$

Also from (15·16), we get

$$\text{Var}(T_1) = \text{Var}(\alpha + \beta T_2) = \beta^2 \text{Var}(T_2)$$

$$\Rightarrow \quad 1 = \beta^2 \quad \Rightarrow \quad \beta = \pm 1 \qquad \text{...[From (15·15)]}$$

But since $\rho(T_1, T_2) = +1$, the coefficient of regression of T_1 and T_2 must be positive.

$$\therefore \quad \beta = 1 \quad \Rightarrow \quad \alpha = 0 \qquad \text{[From 15·17)]}$$

Substituting in (15·16), we get $T_1 = T_2$ as desired.

Theorem 15·4. *Let T_1 and T_2 be unbiased estimators of $\gamma(\theta)$ with efficiencies e_1 and e_2 respectively and $\rho = \rho_\theta$ be the correlation coefficient between them. Then*

$$\sqrt{e_1e_2} - \sqrt{(1-e_1)(1-e_2)} \le \rho \le \sqrt{e_1 e_2} + \sqrt{(1-e_1)(1-e_2)}$$

Proof. Let T be the minimum variance unbiased estimator of $\gamma(\theta)$. Then we are given :

$$E_\theta(T_1) = \gamma(\theta) = E_\theta(T_2), \forall\, \theta \in \Theta \qquad \text{...(15·18)}$$

and

$$e_1 = \frac{V_\theta(T)}{V_\theta(T_1)} = \frac{V}{V_1}, \text{ (say)} \quad \Rightarrow \quad V_1 = \frac{V}{e_1} \qquad \text{...(15·19)}$$

$$e_2 = \frac{V_\theta(T)}{V_\theta(T_2)} = \frac{V}{V_2}, \text{ (say)} \quad \Rightarrow \quad V_2 = \frac{V}{e_2} \qquad \text{...(15·20)}$$

Let us consider another estimator

$$T_3 = \lambda T_1 + \mu T_2 \qquad \text{...(15·21)}$$

which is also unbiased estimator of $\gamma(\theta)$,

i.e., $\quad E(T_3) = (\lambda + \mu)\,\gamma(\theta) = \gamma(\theta)$ [Using (15·18)]

$$\Rightarrow \quad \lambda + \mu = 1 \qquad \text{...(15·22)}$$

$$\begin{aligned} V_\theta(T_3) &= V(\lambda T_1 + \mu T_2) \\ &= \lambda^2 V(T_1) + \mu^2 V(T_2) + 2\lambda\mu \,\text{Cov}(T_1, T_2) \\ &= V\left[\frac{\lambda^2}{e_1} + \frac{\mu^2}{e_2} + 2 \cdot \frac{\lambda\mu\rho}{\sqrt{e_1e_2}}\right] \qquad \text{[Using (15·19) and (15·20)]} \end{aligned}$$

But $\quad V_\theta(T_3) \ge V$, since V is the minimum variance.

$$\therefore \quad \frac{\lambda^2}{e_1} + \frac{\mu^2}{e_2} + \frac{2\rho\lambda\mu}{\sqrt{e_1e_2}} \ge 1 = (\lambda + \mu)^2 \qquad \text{[Using (15·22)]}$$

$$\Rightarrow \quad \left(\frac{1}{e_1} - 1\right)\lambda^2 + \left(\frac{1}{e_2} - 1\right)\mu^2 + 2\lambda\mu\left(\frac{\rho}{\sqrt{e_1e_2}} - 1\right) \ge 0$$

$$\Rightarrow \quad \left(\frac{1}{e_1} - 1\right)\left(\frac{\lambda}{\mu}\right)^2 + 2\left(\frac{\rho}{\sqrt{e_1e_2}} - 1\right)\left(\frac{\lambda}{\mu}\right) + \left(\frac{1}{e_2} - 1\right) \ge 0 \qquad \text{...(15·23)}$$

which is quadratic expression in (λ/μ).

Note that :

$$e_i < 1 \quad \Rightarrow \quad \frac{1}{e_i} > 1 \quad \Rightarrow \quad \left(\frac{1}{e_i} - 1\right) > 0, \; i = 1, 2$$

We know that

$$AX^2 + BX + C \geq 0 \ \forall x, A > 0, C > 0;$$

if and only if

$$\text{Discriminant} = B^2 - 4AC \leq 0 \qquad \ldots(15{\cdot}24)$$

Using (15·24), we get from (15·23) :

$$\left(\frac{\rho}{\sqrt{e_1 e_2}} - 1\right)^2 - \left(\frac{1}{e_1} - 1\right)\left(\frac{1}{e_2} - 1\right) \leq 0$$

$$\Rightarrow \qquad (\rho - \sqrt{e_1 e_2})^2 - (1 - e_1)(1 - e_2) \leq 0$$

$$\Rightarrow \qquad \rho^2 - 2\sqrt{e_1 e_2}\ \rho + (e_1 + e_2 - 1) \leq 0$$

This implies that ρ lies between the roots of the equation

$$\rho^2 - 2\sqrt{e_1 e_2}\ \rho + (e_1 + e_2 - 1) = 0$$

which are given by

$$\frac{1}{2}\left[2\sqrt{e_1 e_2} \pm 2\sqrt{e_1 e_2 - (e_1 + e_2 - 1)}\right]$$

$$= \sqrt{e_1 e_2} \pm \sqrt{(e_1 - 1)(e_2 - 1)}$$

Hence we have :

$$\sqrt{e_1 e_2} - \sqrt{(e_1 - 1)(e_2 - 1)} \leq \rho \leq \sqrt{e_1 e_2} + \sqrt{(e_1 - 1)(e_2 - 1)}$$

$$\Rightarrow \quad \sqrt{e_1 e_2} - \sqrt{(1 - e_1)(1 - e_2)} \leq \rho \leq \sqrt{e_1 e_2} + \sqrt{(1 - e_1)(1 - e_2)} \qquad \ldots(15{\cdot}25)$$

Corollary. If we take $e_1 = 1$ and $e_2 = e$ in (15·25), we get

$$\sqrt{e} \leq \rho \leq \sqrt{e} \quad \Rightarrow \quad \rho = \sqrt{e}$$

This leads to the following important result, which we state in the form of a theorem.

Theorem 15·5. *If T_1 is an MVU estimator of $\gamma(\theta)$, $\theta \in \Theta$ and T_2 is any other unbiased estimator of $\gamma(\theta)$ with efficiency $e = e_\theta$, then the correlation coefficient between T_1 and T_2 is given by*

$$\rho = \sqrt{e} \quad i.e., \quad \rho_\theta = \sqrt{e_\theta}, \forall\, \theta \in \Theta.$$

For an alternate proof, see Examples 15·9 and 15·10.

Theorem 15·6. *If T_1 is an MVUE of $\gamma(\theta)$ and T_2 is any other unbiased estimator of $\gamma(\theta)$ with efficiency $e < 1$, then no unbiased linear combination of T_1 and T_2 can be an MVUE of $\gamma(\theta)$.*

Proof. A linear combination

$$T = l_1 T_1 + l_2 T_2 \qquad \ldots(15{\cdot}27)$$

will be unbiased estimator of $\gamma(\theta)$ if

$$E(T) = l_1 E(T_1) + l_2 E(T_2) = \gamma(\theta), \text{ for all } \theta \in \Theta$$

$$\Rightarrow \qquad l_1 + l_2 = 1 \qquad \ldots(15{\cdot}27a)$$

since we are given $E(T_1) = E(T_2) = \gamma(\theta)$.

We have

$$e = \frac{\text{Var}(T_1)}{\text{Var}(T_2)} \Rightarrow \text{Var}\, T_2 = \frac{\text{Var}\, T_1}{e} \qquad \ldots(15\cdot28)$$

and $\rho = \rho(T_1, T_2) = \sqrt{e}$ [c.f. Theorem 15·5]

From (15·27), on using (15·28), we get

$$\begin{aligned}
\text{Var}\, T &= l_1^2\, \text{Var}(T_1) + l_2^2\, \text{Var}(T_2) + 2l_1 l_2\, \text{Cov}(T_1, T_2) \\
&= l_1^2\, \text{Var}(T_1) + l_2^2\, \text{Var}(T_2) + 2l_1 l_2 \rho \sqrt{\text{Var}(T_1)\, \text{Var}(T_2)} \\
&= \text{Var}(T_1) \left[l_1^2 + \frac{l_2^2}{e} + 2l_1 l_2 \frac{\rho}{\sqrt{e}} \right] \\
&= \text{Var}(T_1) \left[l_1^2 + 2l_1 l_2 + \frac{l_2^2}{e} \right] \qquad (\because \rho \sqrt{e}) \\
&\geq \text{Var}\, T_1\, [l_1^2 + 2l_1 l_2 + l_2^2] \qquad \left(\because 0 < e \leq 1 \Rightarrow \frac{1}{e} \geq 1 \right) \\
&= \text{Var}\, T_1\, (l_1 + l_2)^2 \\
&= \text{Var}(T_1) \qquad \text{[From (15·27}a\text{)]}
\end{aligned}$$

$\Rightarrow$ T cannot be an *MVU* estimator.

Example 15·9. *If T_1 and T_2 be two unbiased estimators of $\gamma(\theta)$ with variances σ_1^2, σ_2^2 and correlation ρ, what is the best unbiased linear combination of T_1 and T_2 and what is the variance of such a combination ?*

[*Delhi Univ.B.Sc. (Stat. Hons.), 1990*]

Solution. Let T_1 and T_2 be two unbiased estimators of $\gamma(\theta)$.

$$\therefore \qquad E(T_1) = E(T_2) = \gamma(\theta) \qquad \ldots(1)$$

Let T be a linear combination of T_1 and T_2 given by

$$T = l_1 T_1 + l_2 T_2 \qquad (*)$$

where l_1, l_2 are arbitrary constants.

$$E(T) = l_1 E(T_1) + l_2 E(T_2) = (l_1 + l_2)\, \gamma(\theta) \qquad \text{[From (1)]}$$

$\therefore$ T is also an unbiased estimator of $\gamma(\theta)$ if and only if

$$l_1 + l_2 = 1 \qquad \ldots(2)$$

Now
$$\begin{aligned}
V(T) &= V(l_1 T_1 + l_2 T_2) \\
&= l_1^2 V(T_1) + l_2^2 V(T_2) + 2l_1 l_2\, \text{Cov}(T_1, T_2) \\
&= l_1^2 \sigma_1^2 + l_2^2 \sigma_2^2 + 2l_1 l_2 \rho\, \sigma_1 \sigma_2 \qquad \ldots(3)
\end{aligned}$$

We want the minimum value of (3) for variations in l_1 and l_2, subject to the condition (2).

$$\therefore \quad \frac{\partial}{\partial l_1} V(T) = 0 = l_1 \sigma_1^2 + l_2 \rho\, \sigma_1 \sigma_2$$

$$\frac{\partial}{\partial l_2} V(T) = 0 \;= l_2\,\sigma_2^2 + l_1\,\rho\,\sigma_1\sigma_2$$

Substracting, we get

$$l_1(\sigma_1^2 - \rho\sigma_1\sigma_2) \;= l_2\,(\sigma_2^2 - \rho\sigma_1\sigma_2)$$

$$\Rightarrow \frac{l_1}{\sigma_2^2 - \rho\sigma_1\sigma_2} = \frac{l_2}{\sigma_1^2 - \rho\sigma_1\sigma_2} = \frac{l_1 + l_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$= \frac{1}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \qquad \text{[From (2)]}$$

$$\therefore \quad l_1 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \text{ and } l_2 = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \qquad \ldots(4)$$

With these values of l_1 and l_2, T given by (*) is the best unbiased linear combination of T_1 and T_2 and its variance is given by (3).

Example 15·10. *Suppose T_1 in the above example is an unbiased minimum variance estimate and T_2 is any other unbiased estimate with variance σ^2/e. Then prove that the correlation between T_1 and T_2 is $\sqrt{e}$.*

Solution. The coefficients of the best linear unbiased combination of T_1 and T_2, given by (*) in Example 15·9 are given by (4).

We are given that $\sigma_1^2 = V(T_1) = \sigma^2$

and $\quad e = \dfrac{V(T_1)}{V(T_2)} = \dfrac{\sigma^2}{V(T_2)} \quad \Rightarrow \quad V(T_2) = \sigma_2^2 = \sigma^2/e$

Substituting in (4) of Example 15·9, we get

$$\left.\begin{aligned} l_1 &= \frac{1-\rho\sqrt{e}}{D} \\ l_2 &= \frac{e-\rho\sqrt{e}}{D} \end{aligned}\right\}, \quad \text{where } D = 1 + e - 2\rho\sqrt{e} \qquad \ldots(5)$$

Hence from (*), the unbiased statistic is

$$T = \frac{[(1-\rho\sqrt{e})\,T_1 + (e-\rho\sqrt{e})\,T_2]}{D}$$

and from (3) the minimum variance is :

$$V(T) = \frac{1}{D^2}\left[(1-\rho\sqrt{e})^2\sigma^2 + (e-\rho\sqrt{e})^2\frac{\sigma^2}{e} + 2(1-\rho\sqrt{e})\,(e-\rho\sqrt{e}).\,\rho.\sigma.\sigma/\sqrt{e}\right]$$

$$= \frac{\sigma^2}{D^2}\left[(1+\rho^2 e - 2\rho\sqrt{e}) + \frac{1}{e}(e^2 + \rho^2 e - 2\rho e\sqrt{e}) + 2\,(1-\rho\sqrt{e})\,(\sqrt{e}-\rho)\rho\right]$$

$$= \frac{\sigma^2}{D^2}\left[1 + \rho^2 e - 2\rho\sqrt{e} + e + \rho^2 - 2\rho\sqrt{e} + 2\,(\rho\sqrt{e} - \rho^2 e - \rho^2 + \rho^3\sqrt{e})\right]$$

$$= \frac{\sigma^2}{D^2}\left[\ 1 - \rho^2 e + e - \rho^2 - 2\rho\ \sqrt{e} + 2\rho^3\sqrt{e}\ \right]$$

$$= \frac{\sigma^2}{D^2}\left[\ (1 + e - 2\rho\sqrt{e}) - \rho^2(e + 1 - 2\rho\ \sqrt{e})\ \right]$$

$$= \frac{\sigma^2(1 - \rho^2)(1 + e - 2\rho\ \sqrt{e})}{(1 + e - 2\rho\ \sqrt{e})^2} = \frac{\sigma^2(1 - \rho^2)}{1 + e - 2\rho\ \sqrt{e}}$$

$$= \frac{\sigma^2\ (1 - \rho^2)}{(1 - \rho^2) + (\sqrt{e} - \rho)^2}$$

$$\therefore \qquad \frac{V(T)}{\sigma^2} = \frac{1 - \rho^2}{(1 - \rho^2) + (\sqrt{e} - \rho)^2} \leq 1 \qquad \ldots(6)$$

Since T_1 is the most efficient estomator,

$$V(T) \not< \sigma^2 \ \Rightarrow\ \frac{V(T)}{\sigma^2} \geq 1 \qquad \ldots(7)$$

From (6) and (7), we get

$$\frac{V(T)}{\sigma^2} = 1,\ \ i.e.,\ \ \frac{1 - \rho^2}{(1 - \rho^2) + (\sqrt{e} - \rho)^2} = 1$$

$$\Rightarrow \qquad (\sqrt{e} - \rho)^2 = 0 \ \Rightarrow\ \rho = \sqrt{e}$$

Aliter. *From (5) onwards.* Since T_1 is given to be the most efficient estimator, it cannot be improved upon (*c.f.* Theorem 15·6). Hence, in order that T defined in (*) is minimum variance unbiased estimator we must have

$$\left.\begin{array}{l} l_1 = 1 \\ l_2 = 0 \end{array}\right\} \Rightarrow\ \rho = \sqrt{e} \qquad \ldots\text{[From (5)]}$$

Remark. This problem leads to the following very important result :

"The correlation coefficient between a most efficient estimator and any other estimator with efficiency e is $\sqrt{e}$*."*

Example 15·11. (*a*) *Show that if a most efficient estimator A and a less efficient estimator B with efficiency e tend to joint normality for large samples, B – A tends to zero correlation with A.*

[***Delhi Univ. B.Sc. (Stat. Hons.), 1988***]

(*b*) *Show that the error in B may be regarded as composed (for large samples) of two parts which are independent, the error in A and the error in (B – A).*

(*c*) *Show further that*

$$V(A - B) = \left(\frac{1}{e} - 1\right) V(A)$$

Solution. (*a*) We have to prove that

$$r[A, (B - A)] = 0 \ \Rightarrow\ \text{Cov}\ (A, B - A) = 0$$

$$\text{Cov}\,[A, (B-A)] = \text{Cov}\,(A, B) - V(A) = \rho\sigma_A\sigma_B - \sigma_A^2,$$

where ρ is the correlation coefficient between A and B.

If we take $\sigma_A = \sigma$, then $\sigma_B = \dfrac{\sigma}{\sqrt{e}}$ and $\rho = \sqrt{e}$ (*c.f.* Theorem 15·5)

$$\therefore \qquad \text{Cov}\,(A, B-A) = \sqrt{e}\,.\,\sigma.\frac{\sigma}{\sqrt{e}} - \sigma^2 = 0$$

Hence $(B-A)$ has zero correlation with A.

(*b*) We have $B = A + (B-A)$

$$\therefore \quad V(B) = V[A + (B-A)] = V(A) + V(B-A) + 2\,\text{Cov}\,(A, B-A)$$

$$= V(A) + V(B-A) \qquad \text{[Using part } (a)]$$

$\Rightarrow$ Error in B = Error in A + Error in $(B-A)$

and since A and $(B-A)$ are independent, [*c.f.* part (*a*) *viz.*, $r(A, B-A) = 0$ and A and B tend to joint normality], the result follows.

(*c*)
$$V(A-B) = V(A) + V(B) - 2\,\text{Cov}\,(A, B)$$

$$= \sigma_A^2 + \sigma_B^2 - 2\,\rho.\sigma_A\,\sigma_B$$

$$= \sigma^2 + \frac{\sigma^2}{e} - 2\sqrt{e}.\,\sigma\,.\frac{\sigma}{\sqrt{e}}$$

$$= \frac{\sigma^2}{e} - \sigma^2 = \left(\frac{1}{e} - 1\right)\sigma^2$$

Example 15·12. *If T_1 and T_2 are two unbiased estimators of $\gamma(\theta)$, having the same variance and ρ is the correlation between them, then show that $\rho \geq 2e - 1$, where e is the efficiency of each estimator.*

Solution. Let T be MVUE of $\gamma(\theta)$. Then, since $V(T_1) = V(T_2)$, the efficiency e of each estimator is given by :

$$e = \frac{V(T)}{V(T_1)} = \frac{V(T)}{V(T_2)} \qquad \ldots(*)$$

Consider another unbiased estimator of $\gamma(\theta)$ *viz.*,

$$T_3 = \frac{1}{2}(T_1 + T_2)$$

$$\Rightarrow \quad V(T_3) = \frac{1}{4}[V(T_1) + V(T_2) + 2\,\text{Cov}\,(T_1, T_2)]$$

$$= \frac{1}{4}\left[\frac{V(T)}{e} + \frac{V(T)}{e} + 2\rho\sqrt{\frac{V(T)}{e}.\frac{V(T)}{e}}\right]$$

$$= \frac{V(T)}{4e}[1 + 1 + 2\rho] = \frac{(1+\rho)\,V(T)}{2e}$$

Since $V(T)$ is the minimum variance,

$$V(T_3) = \frac{(1+\rho)\,.\,V(T)}{2e} \geq V(T)$$

$\Rightarrow \quad 1 + \rho \geq 2e \quad \Rightarrow \quad \rho \geq (2e - 1).$

Aliter. Deduction From (15·25). If T_1 and T_2 have same variances/efficiencies *i.e.*, $e_1 = e_2 = e$, (say) then (15·25) gives

$$e - (1 - e) \leq \rho \leq e + (1 - e) \Rightarrow \rho \geq 2e - 1.$$

15·6. Sufficiency. An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter. More precisely, if $T = t(x_1, x_2, \ldots, x_n)$ is an estimator of a parameter θ, based on a sample $x_1, x_2, \ldots, x_n$ of size n from the population with density $f(x, \theta)$ such that the conditional distribution of $x_1, x_2, \ldots, x_n$ given T, is independent of θ, then T is sufficient estimator for θ.

Illustration. Let $x_1, x_2, \ldots, x_n$ be a random sample from a Bernoulli population with parameter 'p', $0 < p < 1$, *i.e.*,

$$x_i = \begin{cases} 1, \text{ with probability } p \\ 0, \text{ with probability } q = (1 - p) \end{cases}$$

Then $\qquad T = t(x_1, x_2, \ldots, x_n) = x_1 + x_2 + \ldots + x_n \sim B(n, p)$

$$\therefore \qquad P(T = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

The conditional distribution of $(x_1, x_2, \ldots, x_n)$ given T is

$$P[x_1 \cap x_2 \cap \ldots \cap x_n \mid T = k] = \frac{P[x_1 \cap x_2 \cap \ldots \cap x_n \cap T = k]}{P(T = k)}$$

$$= \begin{cases} \dfrac{p^k (1 - p)^{n-k}}{\binom{n}{k} p^k (1 - p)^{n-k}} = \dfrac{1}{\binom{n}{k}} \\ 0, \text{ if } \sum\limits_{i=1}^{n} x_i \neq k \end{cases}$$

Since this does not depend on 'p', $T = \sum\limits_{i=1}^{n} x_i$, is sufficient for '$p$'.

Theorem 15·7. Factorization Theorem (Neyman). The necessary and sufficient condition for a distribution to admit sufficient statistic is provided by the 'factorization theorem' due to Neyman.

Statement *$T = t(x)$ is sufficient for θ if and only if the joint density function L (say), of the sample values can be expressed in the form*

$$L = g_\theta[t(x)].h(x) \qquad \ldots(15·29)$$

where (as indicated) $g_\theta[t(x)]$ depends on θ and x only through the value of $t(x)$ and $h(x)$ is independent of θ.

Remarks 1. It should be clearly understood that by 'a function independent of θ' we not only mean that it does not involve θ but also that its domain does not contain θ. For example, the function

$$f(x) = \frac{1}{2a}, \; a - \theta < x < a + \theta \; ; \; -\infty < \theta < \infty$$

depends on θ.

2. It should be noted that the original sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, is always a sufficient statistic.

3. The most general form of the distributions admitting sufficient statistic is *Koopman's form* and is given by

$$L = L(\mathbf{x}, \theta) = g(\mathbf{x}).h(\theta).\exp\{a(\theta)\psi(\mathbf{x})\} \qquad \ldots(15\text{·}30)$$

where $h(\theta)$ and $a(\theta)$ are functions of the parameter θ only and $g(\mathbf{x})$ and $\psi(\mathbf{x})$ are the functions of the sample observations only.

Equation (15·30) represents the famous *exponential family of distributions*, of which most of the common distributions like the binomial, the Poisson and the normal with unknown mean and variance, are the members.

4. Invariance Property of Sufficient Estimator.

If T is a sufficient estimator for the parameter θ and if $\psi(T)$ is a one to one function of T, then $\psi(T)$ is sufficient for $\psi(\theta)$.

5. Fisher-Neyman Criterion. *A statistic $t_1 = t_1(x_1, x_2, \ldots, x_n)$ is sufficient estimator of parameter θ if and only if the likelihood function (joint p.d.f. of the sample) can be expressed as :*

$$L = \prod_{i=1}^{n} f(x_i, \theta)$$

$$= g_1(t_1, \theta).\, k(x_1, x_2, \ldots, x_n) \qquad \ldots(15\text{·}31)$$

where $g_1(t_1, \theta)$ is the p.d.f. of statistic t_1 and $k(x_1, x_2, \ldots, x_n)$ is a function of sample observations only independent of θ.

Note that this method requires the working out of the p.d.f. (p.m.f.) of the statistic $t_1 = t(x_1, x_2, \ldots, x_n)$, which is not always easy.

Example 15·13. *Let $x_1, x_2, \ldots, x_n$ be a random sample from a uniform population on* $[0, \theta]$. *Find a sufficient estimator for θ.*

[*Madras Univ. B.Sc., Oct. 1992*]

Solution. We are given

$$f_\theta(x_i) = \begin{cases} \frac{1}{\theta}, & 0 \le x_i \le \theta \\ 0, & \text{otherwise} \end{cases}$$

Let
$$k(a, b) = \begin{cases} 1, & \text{if } a \le b \\ 0, & \text{if } a > b \end{cases}$$

Then
$$f_\theta(x_i) = \frac{k(0, x_i)\, k(x_i, \theta)}{\theta},$$

$$L = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \left[\frac{k(0, x_i)\, k(x_i, \theta)}{\theta}\right]$$

$$= \frac{k(0, \min_{1 \le i \le n} x_i) \cdot k(\max_{1 \le i \le n} x_i, \theta)}{\theta^n} = g_\theta[t(\mathbf{x})]\, h(\mathbf{x})$$

where $g_\theta[t(\mathbf{x})] = \dfrac{k\{t(\mathbf{x}), \theta\}}{\theta^n}$, $t(\mathbf{x}) = \max\limits_{1 \le i \le n} x_i$ and $h(\mathbf{x}) = k(0, \min\limits_{1 \le i \le n} x_i)$

Hence by Factorization Theorem, $T = \max\limits_{1 \le i \le n} x_i$, is sufficient statistic for θ.

Aliter. We have $L = \prod\limits_{i=1}^{n} f(x_i, \theta) = \dfrac{1}{\theta^n}$; $0 < x_i < \theta$ ...(*i*)

If $t = \max(x_1, x_2, \ldots, x_n) = x_{(n)}$, then p.d.f. of T is given by :

$$g(t, \theta) = n\,[F(x_{(n)})]^{n-1} \cdot f(x_{(n)}) \qquad \ldots(ii)$$

We have $F(x) = P(X \le x) = \displaystyle\int_0^x f(x, \theta)\, dx = \int_0^x \frac{1}{\theta} \cdot dx = \frac{x}{\theta}$

$$\therefore \quad g(t, \theta) = n\left[\frac{x_{(n)}}{\theta}\right]^{n-1}\left(\frac{1}{\theta}\right) \qquad \text{[From (ii)]}$$

$$= \frac{n}{\theta^n}\,[x_{(n)}]^{n-1}$$

Rewriting (*i*), we get

$$L = \frac{n\,[x_{(n)}]^{n-1}}{\theta^n} \cdot \frac{1}{n\,[x_{(n)}]^{n-1}}$$

$$= g(t, \theta) \cdot h(x_1, x_2, \ldots, x_n)$$

Hence by Fisher-Neyman criterion, the statistic $t = x_{(n)}$, is sufficient estimator for θ.

Example 15·14. *Let $x_1, x_2, \ldots, x_n$ be a random sample from $N(\mu, \sigma^2)$ population. Find sufficient estimators for μ and σ^2.*

Solution. Let us write

$$\theta = (\mu, \sigma^2)\ ; -\infty < \mu < \infty,\ 0 < \sigma^2 < \infty$$

Then

$$L = \prod_{i=1}^{n} f_\theta(x_i) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum x_i + n\mu^2\right)\right]$$

$$= g_\theta[t(\mathbf{x})] \cdot h(\mathbf{x})$$

where

$$g_\theta[t(\mathbf{x})] = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\{t_2(\mathbf{x}) - 2\mu t_1(\mathbf{x}) + n\mu^2\}\right],$$

$$t(\mathbf{x}) = [t_1(\mathbf{x}), t_2(\mathbf{x})] = (\Sigma x_i, \Sigma x_i^2) \text{ and } h(\mathbf{x}) = 1$$

Thus $t(\mathbf{x}) = \Sigma x_i$ is sufficient for μ and $t_2(\mathbf{x}) = \Sigma x_i^2$, is sufficient for σ^2.

Example 15·15. *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f.*

$$f(x, \theta) = e^{-(x-\theta)}, \theta < x < \infty \; ; -\infty < \theta < \infty$$

Obtain sufficient statistic for θ.

Solution. Here

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \prod_{i=1}^{n} \left[e^{-(x_i - \theta)}\right]$$

$$= \exp\left[-\sum_{i=1}^{n} x_i\right] \times \exp(n\theta) \qquad \ldots(*)$$

Let $Y_1, Y_2, \ldots, Y_n$ denote the order statistics of the random sample such that $Y_1 < Y_2 < \ldots < Y_n$. The p.d.f. of the smallest observation Y_1 is given by

$$g_1(y_1, \theta) = n[1 - F(y_1)]^{n-1} f(y_1, \theta)$$

where $F(\cdot)$ is the distribution function corresponding to p.d.f. $f(\cdot)$.

Now $\quad F(x) = \int_{\theta}^{x} e^{-(x-\theta)} \, dx = \left| \dfrac{e^{-(x-\theta)}}{-1} \right|_{\theta}^{x} = 1 - e^{-(x-\theta)}$

$$\therefore \; g_1(y_1, \theta) = n\left[e^{-(y_1 - \theta)}\right]^{n-1} . e^{-(y_1 - \theta)}$$

$$= n\, e^{-n(y_1 - \theta)}, \theta < y_1 < \infty$$

$$= 0, \text{ otherwise}$$

Thus the likelihood function of $X_1, X_2, \ldots, X_n$ may be expressed as

$$L = e^{n\theta} \exp\left(-\sum_{i=1}^{n} x_i\right)$$

$$= n \exp\{-n(y_1 - \theta)\} \left[\frac{\exp\left(-\sum_{i=1}^{n} x_i\right)}{n \exp(-n y_1)}\right]$$

$$= g_1(\min x_i, \theta) \left[\frac{\exp\left(-\sum_{i=1}^{n} x_i\right)}{n \exp(-n \min x_i)}\right]$$

Hence by Fisher-Neyman criterion, the first order statistic

$Y_1 = \min(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for θ.

Example 15·16. *Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with p.d.f.*

$$f(x, \theta) = \theta x^{\theta - 1} \; ; 0 < x < 1, \; \theta > 0.$$

Show that $t_1 = \prod_{i=1}^{n} X_i$, *is sufficient for* θ.

[**Delhi Univ. B.Sc. (Stat. Hons.), 1988; Agra Univ. B.Sc., 1992**]

Solution. $L(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \theta^n \prod_{i=1}^{n} (x_i^{\theta-1})$

$$= \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta} \cdot \frac{1}{\left(\prod_{i=1}^{n} x_i\right)}$$

$$= g(t_1, \theta) . h(x_1, x_2, \ldots, x_n), \text{ (say).}$$

Hence by Factorisation Theorem,

$$t_1 = \prod_{i=1}^{n} X_i, \text{ is sufficient estimator for } \theta.$$

Example 15·17. *Let $X_1, X_2, \ldots, X_n$ be a random sample from Cauchy population :*

$$f(x, \theta) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}\ ; -\infty < x < \infty, -\infty < \theta < \infty.$$

Examine if there exists a sufficient statistic for θ.

Solution. $L(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \frac{1}{\pi^n} \cdot \prod_{i=1}^{n} \left[\frac{1}{1 + (x_i - \theta)^2}\right]$

$$\neq g(t_1, \theta) . h(x_1, x_2, \ldots, x_n).$$

Hence by Factorisation Theorem, there is no single statistic, which alone, is sufficient estimator of θ.

However,

$$L(\mathbf{x}, \theta) = k_1(X_1, X_2, \ldots, X_n, \theta).\ k_2(X_1, X_2, \ldots, X_n)$$

$\Rightarrow$ The whole set $(X_1, X_2, \ldots, X_n)$ is jointly sufficient for θ.

15·7. Cramer-Rao Inequality

Theorem 15·8. *If t is an unbiased estimator for $\gamma(\theta)$, a function of parameter θ, then*

$$Var(t) \geq \frac{\left[\frac{d}{d\theta} \cdot \gamma(\theta)\right]^2}{E\left[\frac{\partial}{\partial\theta} \log L\right]^2} = \frac{[\gamma'(\theta)]^2}{I(\theta)} \quad \ldots(15·32)$$

where $I(\theta)$ is the information on θ, supplied by the sample.

In other words, Cramer-Rao inequality provides a lower bound $[\gamma'(\theta)]^2/I(\theta)$, to the variance of an unbiased estimator of $\gamma(\theta)$.

Proof. In proving this result, we assume that there is only a single parameter θ which is unknown. We also take the case of continuous *r.v.* The case of descrete random variables can be dealt with similarly on replacing the multiple integrals by appropriate multiple sums.

We further make the following assumptions, which are known as the *Regularity conditions* for *Cramer-Rao Inequality*.

(1) The parameter space Θ is a non-degenerate open interval on the real line R^1 $(-\infty, \infty)$.

(2) For almost all $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and for all $\theta \in \Theta$, $\dfrac{\partial}{\partial \theta} L(\mathbf{x}, \theta)$ exists, the exceptional set, if any, is independent of θ.

(3) The range of integration is independent of the parameter θ, so that $f(x, \theta)$ is differentiable under integral sign.

If range is not independent of θ and f is zero at the extremes of the range, *i.e.*, $f(a, \theta) = 0 = f(b, \theta)$, then

$$\frac{\partial}{\partial \theta} \int_a^b f dx = \int_a^b \frac{\partial f}{\partial \theta} dx - f(a, \theta) \frac{\partial a}{\partial \theta} + f(b, \theta) \frac{\partial b}{\partial \theta}$$

$$\Rightarrow \qquad \frac{\partial}{\partial \theta} \int_a^b f dx = \int_a^b \frac{\partial f}{\partial \theta} dx, \text{ since } f(a, \theta) = 0 = f(b, \theta)$$

(4) The conditions of uniform convergence of integrals are satisfied so that differentiation under the integral sign is valid.

(5) $I(\theta) = E\left[\left\{\dfrac{\partial}{\partial \theta} \log L(x, \theta)\right\}^2\right]$, exists and is positive for all $\theta \in \Theta$.

Let X be a *r.v.* following the p.d.f. $f(x, \theta)$ and let L be the likelihood function of the random sample $(x_1, x_2, \ldots, x_n)$ from this population. Then

$$L = L(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Since L is the joint p.d.f. of $(x_1, x_2, \ldots, x_n)$,

$$\int L(\mathbf{x}, \theta)\, d\mathbf{x} = 1,$$

where $$\int d\mathbf{x} = \iint \ldots \int dx_1\, dx_2 \ldots dx_n.$$

Differentiating w.r. to θ and using regularity conditions given above, we get :

$$\int \frac{\partial}{\partial \theta} L\, d\mathbf{x} = 0 \quad \Rightarrow \quad \int \left(\frac{\partial}{\partial \theta} \log L\right) L\, d\mathbf{x} = 0$$

$$\Rightarrow \qquad E\left(\frac{\partial}{\partial \theta} \log L\right) = 0 \qquad \ldots(15{\cdot}33)$$

Let $t = t(x_1, x_2, \ldots, x_n)$ be an unbiased estimator of $\gamma(\theta)$ such that

$$E(t) = \gamma(\theta) \quad \Rightarrow \quad \int t \,.\, L\, d\mathbf{x} = \gamma(\theta) \qquad \ldots(15{\cdot}34)$$

Differentiating w.r. to θ, we get

$$\int t \,.\, \frac{\partial L}{\partial \theta} d\mathbf{x} = \gamma'(\theta) \quad \Rightarrow \quad \int t \left(\frac{\partial}{\partial \theta} \log L\right) L\, d\mathbf{x} = \gamma'(\theta)$$

$$\Rightarrow \qquad E\left(t \cdot \frac{\partial}{\partial \theta} \log L\right) = \gamma'(\theta) \qquad \ldots(15 \cdot 35)$$

$$\text{Cov}\left(t, \frac{\partial}{\partial \theta} \log L\right) = E\left[t \cdot \frac{\partial}{\partial \theta} \log L\right] - E(t) \cdot E\left(\frac{\partial}{\partial \theta} \log L\right)$$

$$= \gamma'(\theta) \qquad \text{[From (15·33) and (15·35)]}$$

We have :

$$[r(X, Y)]^2 \leq 1 \quad \Rightarrow \quad [\text{Cov}(X, Y)]^2 \leq \text{Var}(X). \text{Var}(Y)$$

$$\therefore \qquad \left[\text{Cov}\left(t, \frac{\partial}{\partial \theta} \log L\right)\right]^2 \leq \text{Var}\, t. \ \text{Var}\left(\frac{\partial}{\partial \theta} \log L\right)$$

$$\Rightarrow \qquad [\gamma'(\theta)]^2 \leq \text{Var}\, t\left[E\left\{\frac{\partial}{\partial \theta} \log L\right\}^2 - \left\{E\left(\frac{\partial}{\partial \theta} \log L\right)\right\}^2\right]$$

$$\Rightarrow \qquad [\gamma'(\theta)]^2 \leq \text{Var}\, t \cdot E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right] \qquad \text{[Using (15·33)]} \qquad \ldots(15 \cdot 36)$$

$$\Rightarrow \qquad \text{Var}(t) \geq \frac{[\gamma'(\theta)]^2}{E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right]} \qquad \ldots(15 \cdot 36a)$$

which is Cramer-Rao Inequality.

Corollary. If t is an unbiased estimator of parameter θ *i.e.,*

$$E(t) = \theta \quad \Rightarrow \quad \gamma(\theta) = \theta \quad \Rightarrow \quad \gamma'(\theta) = 1,$$

then from (15·36*a*), we get

$$\text{Var}(t) \geq \frac{1}{E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right]} = \frac{1}{I(\theta)} \qquad \ldots(15 \cdot 37)$$

where

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right] \qquad \ldots(15 \cdot 37a)$$

is called by R.A. Fisher as the *amount of information* on θ supplied by the sample $(x_1, x_2, \ldots, x_n)$ and its reciprocal $1/I(\theta)$, as the *information limit* to the variance of estimator $t = t(x_1, x_2, \ldots, x_n)$.

Remarks. 1. An unbiased estimator t of $\gamma(\theta)$ for which Cramer-Rao lower bound in (15·32) is attained is called a *minimum variance bound* (*MVB*) *estimator.*

2. We have :

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log L\right)^2\right] = -E\left[\frac{\partial^2}{\partial \theta^2} \log L\right] \qquad \ldots(15 \cdot 38)$$

and

$$I(\theta) = n\left[\frac{\partial}{\partial \theta} \log f(x, \theta)\right]^2 = -n\left[\frac{\partial^2}{\partial \theta^2} \log f\right] \qquad \ldots(15 \cdot 38a)$$

Proof. We have proved in (15·33),

$$E\left(\frac{\partial}{\partial\theta}\log L\right)=0 \qquad \ldots(*)$$

Also

$$\left(\frac{\partial^2}{\partial\theta^2}\log L\right)L = \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log L\right).L\right]-\left(\frac{\partial}{\partial\theta}\log L\right).\frac{\partial L}{\partial\theta}$$

$$= \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log L\right)L\right]-\left(\frac{\partial}{\partial\theta}\log L\right)^2.L$$

Integrating both sides w.r. to $\mathbf{x}=(x_1, x_2, \ldots, x_n)$, we get

$$E\left(\frac{\partial^2}{\partial\theta^2}\log L\right) = \frac{\partial}{\partial\theta}.E\left(\frac{\partial}{\partial\theta}\log L\right)-E\left(\frac{\partial}{\partial\theta}\log L\right)^2$$

$$= -E\left(\frac{\partial}{\partial\theta}\log L\right)^2 \qquad \text{[Using (*)]}$$

$$\Rightarrow \qquad I(\theta) = E\left(\frac{\partial}{\partial\theta}\log L\right)^2 = -E\left(\frac{\partial^2}{\partial\theta^2}\log L\right),$$

a form which is more convenient to use in practice.

$$\text{Also} \qquad I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log L\right)^2\right]=E\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i, \theta)\right]^2$$

$$= E\left[\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\theta}\log f(x_i, \theta)\right\}^2 + \sum_{i\neq j=1}^{n}\left\{\left(\frac{\partial}{\partial\theta}\log f(x_i, \theta)\right)\left(\frac{\partial}{\partial\theta}\log f(x_j, \theta)\right)\right\}\right]$$

$$= n\,.\,E\left[\frac{\partial}{\partial\theta}\log f(x, \theta)\right]^2, \qquad \text{[On using (*)]}$$

since x_i's; $i = 1, 2, \ldots, n$ are i.i.d. $r.v.$'s.

15·7·1. Conditions for the Equality Sign in Cramer-Rao (C.R.) Inequality.

In proving (15·32) we used [*c.f.* (15·36) that

$$[\gamma'(\theta)]^2 \leq E\,[t-\gamma(\theta)]^2\,.\,E\left(\frac{\partial}{\partial\theta}\log L\right)^2 \qquad \ldots(15\cdot39)$$

The sign of equality will hold in C.R. Inequality if and only if the sign of equality holds in (15·39). The sign of equality will hold in (15·39) by Cauchy Schwartz Inequality, if and only if the variables $[t-\gamma(\theta)]$ and $\left(\frac{\partial}{\partial\theta}\log L\right)$ are proportional to each other, *i.e.*,

$$\frac{t-\gamma(\theta)}{\frac{\partial}{\partial\theta}\log L} = \lambda = \lambda(\theta)$$

where λ is a constant independent of $(x_1, x_2, \ldots, x_n)$ but may depend on θ.

$$\therefore \qquad \frac{\partial}{\partial\theta}\log L = \frac{t-\gamma(\theta)}{\lambda(\theta)} = [t-\gamma(\theta)]\,A(\theta) \qquad \ldots(15\cdot40)$$

where $\qquad A = A(\theta) = 1/[\lambda(\theta)]$, say.

Hence *a necessary and sufficient condition for an unbiased estimator t to attain the lower bound of its variance is given by (15·40).*

Further, the C-R minimum variance bound is given by :

$$\text{Var}(t) = [\gamma'(\theta)]^2 \Big/ E\left(\frac{\partial}{\partial\theta}\log L\right)^2 \qquad \ldots(15\cdot41)$$

But $\qquad E\left(\dfrac{\partial}{\partial\theta}\log L\right)^2 = E\left[A(\theta)\,.\,\{t-\gamma(\theta)\}\right]^2 \qquad$ [From (15·40)]

$$= [A(\theta)]^2 \,.\, E\,[t-\gamma(\theta)]^2$$

$$= [A(\theta)]^2 \,.\, \text{Var}(t)$$

Substituting in (15·41), we get

$$\text{Var}(t) = \frac{[\gamma'(\theta)]^2}{[A(\theta)]^2\,.\,\text{Var}(t)}$$

$$\Rightarrow \qquad \text{Var}(t) = \left|\frac{\gamma'(\theta)}{A(\theta)}\right| = |\gamma'(\theta).\lambda(\theta)| \qquad \ldots(15\cdot42)$$

Hence if the likelihood function L is expressible in the form (15·40) viz.,

$$\frac{\partial}{\partial\theta}\log L = \frac{t-\gamma(\theta)}{\lambda(\theta)} = [t-\gamma(\theta)]\,.\,A(\theta),$$

then

(i) t is an unbiased estimator of $\gamma(\theta)$.

(ii) Minimum Variance Bound (MVB) estimator (t) for $\gamma(\theta)$ exists, and

(iii) $\text{Var}(t) = \left|\dfrac{\gamma'(\theta)}{A(\theta)}\right| = |\gamma'(\theta)\,\lambda(\theta)|$

The importance of this result lies in that fact that C.R. inequality, in addition to find if MVBU estimator for $\gamma(\theta)$ exists, also gives us the variance of such an estimator, which is given by (15·42).

Remarks 1. If $\gamma(\theta) = \theta$, *i.e.*, if t is an unbiased estimator of θ, then (15·40) can be written as :

$$\frac{\partial}{\partial\theta}\log L = \frac{t-\theta}{\lambda} \qquad \ldots(15\cdot43)$$

Hence if (15·43) holds, then t is an MVB estimator for θ with

$$\text{Var}(t) = |\lambda(\theta)| = 1/|A(\theta)| \qquad \ldots(15\cdot43a)$$

2. We have seen in (15·40) that an MVB estimator exists for $\gamma(\theta)$ if

$$\frac{\partial}{\partial\theta}\log L = \frac{t-\gamma(\theta)}{\lambda} = [t-\gamma(\theta)]\,.\,\frac{1}{\lambda}, \qquad \ldots(*)$$

where $\lambda = \lambda(\theta)$, say. If we write

$$\int \frac{1}{\lambda} d\theta = \alpha(\theta),$$

then integrating (*) w.r. to θ (by parts), we get

$$\log L = [t - \gamma(\theta)] \alpha(\theta) + \int \alpha(\theta) . \gamma'(\theta) d\theta + k(\mathbf{x})$$

$$\Rightarrow \quad \log L = [t - \gamma(\theta)] \alpha(\theta) + \beta(\theta) + k(\mathbf{x}) \qquad \ldots(15·44)$$

where $\alpha(\theta)$ and $\beta(\theta)$ are arbitrary functions of θ and $k(\mathbf{x}) = k(x_1, x_2, \ldots, x_n)$, is an arbitrary function of x_i's independent of θ.

Hence $\log f(x, \theta) = [t - \gamma(\theta)] A_1(\theta) + B_1(\theta) + k_1(\mathbf{x})$

$$\Rightarrow \quad f(x, \theta) = g(\mathbf{x}) . h(\theta) . \exp [a(\theta) . \psi(\mathbf{x})] \qquad \ldots(15·44a)$$

which is the necessary and sufficient condition for the existence of a sufficient statistic [*c.f.* Koopman's form, Equation (15·30) in Remark 3 to § 15·6)]. Hence *an MVB estimator for $\gamma(\theta)$ exists if and only if there exists a sufficient estimator for $\gamma(\theta)$.*

This suggests that in our search for an MVB estimator for $\gamma(\theta)$, we need to confine ourselves to sufficient estimators of $\gamma(\theta)$ alone.

This explains why the method failed in the case of Cauchy population [*c.f.* Example 15·19], where no sufficient estimator exists and its success in the case of normal population [*c.f.* Example 15·18, where $\bar{x}$ is sufficient for μ and Example 15·20, $\sum_{i=1}^{n} x_i^2/n$ is sufficient for σ^2].

Example 15·18. *Obtain the MVB estimator for μ in the normal population $N(\mu, \sigma^2)$, where σ^2 is known.*

Solution. If $x_1, x_2, \ldots, x_n$ is a random sample of size n from the normal population, then

$$L = \prod_{i=1}^{n} f(x_i, \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n . \exp\left\{-\sum_{i=1}^{n} (x_i - \mu)^2/2\sigma^2\right\}$$

$$\log L = -n \log (\sqrt{2\pi}\,\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$= k - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2,$$

where k is a constant independent of μ, (σ being known).

$$\frac{\partial}{\partial\mu} \log L = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [2(x_i - \mu)(-1)]$$

$$= \frac{\sum_{i=1}^{n} (x_i - \mu)}{\sigma^2} = \frac{\sum x_i - n\mu}{\sigma^2} = \frac{(\bar{x} - \mu)}{\sigma^2/n}$$

which is of the form (15·40).

Hence $\bar{x}$ is an MVB unbiased estimator for μ and $V(\hat{\mu}) = V(\bar{x}) = \dfrac{\sigma^2}{n}$.

Example 15·19. *Find if MVB estimator exists for θ in the Cauchy's population :*

$$dF(x, \theta) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}, -\infty < x < \infty,$$

Solution. Here

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \left(\frac{1}{\pi}\right)^n \prod_{i=1}^{n} \left[\frac{1}{1 + (x_i - \theta)^2}\right]$$

$\therefore$
$$\log L = -n \log \pi - \sum_{i=1}^{n} \log [1 + (x_i - \theta)^2]$$

$\Rightarrow$
$$\frac{\partial}{\partial \theta} \log L = 2 \sum_{i=1}^{n} \left[\frac{(x_i - \theta)}{1 + (x_i - \theta)^2}\right]$$

Since this cannot be expressed in form (15·40), MVB estimator does not exist for θ, in the Cauchy's population and so Cramer Rao lower bound is not attainable by the variance of any unbiased estimator θ.

Example 15·20. *A random sample $x_1, x_2, \ldots, x_n$ is taken from a normal population with mean zero and variance σ^2. Examine if $\sum_{i=1}^{n} x_i^2/n$ is an MVB estimator for σ^2.*

Solution. Since $X \sim N(0, \sigma^2)$,

$$f(x, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right), -\infty < x < \infty$$

$$L = \prod_{i=1}^{n} f(x_i, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n} (x_i^2/2\sigma^2)\right\}$$

$\Rightarrow$
$$\log L = -\frac{n}{2} \log (2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2$$

$$\frac{\partial}{\partial \sigma^2} \log L = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} x_i^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\sigma^2}{2\sigma^4}$$

$$= \frac{\left(\sum_{i=1}^{n} x_i^2/n\right) - \sigma^2}{(2\sigma^4/n)},$$

which is of the form (15·40).

Hence $\hat{\sigma}^2 = \sum_{i=1}^{n} \dfrac{x_i^2}{n}$, is an MVB estimator and $V(\hat{\sigma}^2) = \dfrac{2\sigma^4}{n}$

Example 15·21. *Show that* $\overline{X} = \sum_{i=1}^{n} X_i / n$, *in random sampling from*

$$f(x, \theta) = \begin{cases} (1/\theta) \exp(-x/\theta), & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \qquad \ldots(*)$$

where $0 < \theta < \infty$, *is an MVB estimator of* θ *and has variance* θ^2/n.

Solution. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from population with p.d.f. in (*). Then

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \frac{1}{\theta^n} \cdot \exp\left[- \sum_{i=1}^{n} x_i / \theta \right]$$

$$\Rightarrow \quad \log L = -n \log \theta - \frac{1}{\theta} \cdot \sum_{i=1}^{n} x_i$$

$$\therefore \quad \frac{\partial}{\partial \theta} \log L = -\frac{n}{\theta} + \frac{1}{\theta^2} \cdot \sum_{i=1}^{n} x_i$$

$$\Rightarrow \quad \frac{\partial^2}{\partial \theta^2} \log L = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^{n} x_i$$

$$\therefore \quad I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log L \right] = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \cdot \sum_{i=1}^{n} E(x_i)$$

In sampling from exponential population (*), we have

$$E(X) = \theta \text{ and } \mathrm{Var}(X) = \theta^2 \qquad \ldots(**)$$

$$\therefore \quad I(\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \cdot \sum_{i=1}^{n} (\theta) \qquad (\because x_i\text{'s are i.i.d.})$$

$$= -\frac{n}{\theta^2} + \frac{2}{\theta^3} \cdot n\theta = \frac{n}{\theta^2}$$

Also $\quad \gamma(\theta) = \theta \quad \Rightarrow \quad \gamma'(\theta) = 1.$

Hence Cramer Rao lower bound to the variance of an unbiased estimator of θ is :

$$\frac{[\gamma'(\theta)]^2}{I(\theta)} = \frac{1}{(n/\theta^2)} = \frac{\theta^2}{n} \qquad \ldots(***)$$

Consider the estimator $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

We have :

$$E(\overline{X}) = \frac{1}{n} \sum_{i=1}^{n} E(x_i) = \frac{1}{n} \sum_{i=1}^{n} (\theta) = \theta$$

$\Rightarrow \quad \overline{X}$ is an unbiased estimator of θ.

Also $\quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} = \frac{\mathrm{Var}\, X}{n} = \frac{\theta^2}{n} \qquad$ [From (**)]

Thus we see that Var $(\bar{X})$ coincides with the Cramer-Rao lower bound obtained in (***). Hence $\bar{X}$, the sample mean is an MVB unbiased estimator, for θ.

Aliter. A more convenient way of doing this problem is as follows :

We have

$$\frac{\partial}{\partial\theta}\log L = -\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i - n\theta}{\theta^2}$$

$$= \frac{\bar{X}-\theta}{(\theta^2/n)} = \frac{\bar{X}-\theta}{\lambda(\theta)}, \text{ (say)}$$

which is of the form (15·40).

Hence $\bar{X}$ is an MVB unbiased estimator of θ and Var $(\bar{X}) = \lambda(\theta) = \theta^2/n$.

Example 15·22. *Given the probability density function*

$$f(x:\theta) = [\pi\{1+(x-\theta)^2\}]^{-1}; -\infty < x < \infty, -\infty < \theta < \infty \qquad \text{...(*)}$$

show that the Cramer-Rao lower bound of variance of an unbiased estimator of θ is $\frac{2}{n}$, where n is the size of the random sample from this distribution...

[Sri Venkateswara Univ M.Sc., 1992]

Solution. $\log f = -\log\pi - \log[1+(x-\theta)^2]$

$$\frac{\partial \log f}{\partial\theta} = \frac{2(x-\theta)}{[1+(x-\theta)^2]}$$

$$E\left(\frac{\partial \log f}{\partial\theta}\right)^2 = \int_{-\infty}^{\infty} \frac{4(x-\theta)^2}{[1+(x-\theta)^2]^2} f(x,\theta)\,dx$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{4(x-\theta)^2}{[1+(x-\theta)^2]^3}\,dx$$

Put $x-\theta = \tan\phi \Rightarrow dx = \sec^2\phi\, d\phi$.

$$\therefore E\left(\frac{\partial}{\partial\theta}\log f\right)^2 = \frac{2}{\pi}\int_0^{\pi/2} \frac{4\tan^2\phi}{\sec^6\phi}\sec^2\phi\,d\phi = \frac{2}{\pi}\int_0^{\pi/2}\frac{4\sin^2\phi}{\cos^2\phi}\cos^4\phi\,d\phi$$

$$= \frac{2}{\pi}\int_0^{\pi/2} 4\sin^2\phi\cos^2\phi\,d\phi = \frac{8}{\pi}\int_0^{\pi/2}(\cos^2\phi-\cos^4\phi)\,d\phi$$

$$= \frac{8}{\pi}\left[\frac{1}{2}\cdot\frac{\pi}{2} - \frac{3.1}{4.2}\cdot\frac{\pi}{2}\right]$$

(Using reduction formula for $\int_0^{\pi/2} \cos^n x\,dx$).

$$= \frac{8}{\pi}\left[\frac{\pi}{4} - \frac{3\pi}{16}\right] = \frac{1}{2}$$

Hence Cramer-Rao lower bound is

$$= \frac{1}{nE\left(\frac{\partial \log f}{\partial \theta}\right)^2} = \frac{1}{n\left[\frac{1}{2}\right]} = \frac{2}{n}.$$

Example 15·23. *Prove that under certain general conditions of regularity to be stated clearly the mean square deviation $E(\hat{\theta} - \theta)^2$ of an estimator $\hat{\theta}$ of the parameter θ, can never fall below a positive limit depending only on the density function $f(x, \theta)$, the size of the sample and the bias of the estimate.*

Solution. We have proved Cramer-Rao's inequality

$$V(\hat{\theta}) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}, \text{ where } E(\hat{\theta}) = \psi(\theta) \qquad \ldots(*)$$

Now

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E[\hat{\theta} - \psi(\theta) + \psi(\theta) - \theta]^2 \\ &= E[\hat{\theta} - \psi(\theta)]^2 + [\theta - \psi(\theta)]^2 + 2[\psi(\theta) - \theta]\,.\,E[\hat{\theta} - \psi(\theta)] \\ &= V(\hat{\theta}) + [\theta - \psi(\theta)]^2 \end{aligned}$$

$$\therefore \quad E(\hat{\theta} - \theta)^2 \geq \frac{[\psi'(\theta)]^2}{I(\theta)} + [\theta - \psi(\theta)]^2 \quad \text{[Using (*)]} \qquad \ldots(**)$$

Let $\hat{\theta}$ be a 'biased' estimator of θ with bias given by $b(\theta)$

i.e., $\quad E(\hat{\theta}) = \theta + b(\theta) = \psi(\theta)$, (say).

$\therefore \quad \psi(\theta) - \theta = b(\theta)$

From (**), we get

$$\therefore \quad E(\hat{\theta} - \theta)^2 \geq \frac{\left[1 + \frac{\partial}{\partial \theta} b(\theta)\right]^2}{I(\theta)} + [b(\theta)]^2 > 0,$$

where $\quad I(\theta) = n \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \theta} \log f\right)^2 f(x, \theta)\, dx > 0$

This proves the result.

15·8. Complete Family of Distributions. Consider a statistic $T = (x_1, x_2, \ldots, x_n)$, based on a random sample of size n from the population $f(x, \theta)$, $\theta \in \Theta$. The distribution of the statistic T will, in general, depend on θ. Hence corresponding to T, we again have a family of distributions, say, $\{g(t, \theta), \theta \in \Theta\}$.

Definition. The statistic $T = t(x)$, or more precisely the family of distributions $\{g(t, \theta), \theta \in \Theta\}$ is said to be complete for θ if

$$E_\theta[h(T)] = 0 \text{ for all } \theta \Rightarrow P_\theta[h(T) = 0] = 1 \qquad \ldots(15\cdot45)$$

$$\left.\begin{array}{ll} \text{i.e.,} & \int h(t)\, g(t, \theta)\, dt = 0 \text{ for all } \theta \in \Theta \\ \text{or} & \sum_t h(t)\, g(t, \theta) = 0 \text{ for all } \theta \in \Theta \end{array}\right\} \quad \ldots(15\cdot45a)$$

$\Rightarrow \quad h(T) = 0$, for all $\theta \in \Theta$, almost surely (*a.s.*). ...(15·45*b*)

The concept of complete sufficient statistic is specially useful in Rao-Blackwell Theorem [*c.f.* § 15·9].

Example 15·24. *Let $X_1, X_2, \ldots, X_n$ be a random sample from Bernoulli distribution :*

$$f(x, \theta) = \begin{cases} \theta^x (1-\theta)^{1-x} \; ; x = 0, 1 \\ 0 \quad , \textit{otherwise} \end{cases}$$

Show that $\sum_{i=1}^{n} X_i$, is a complete sufficient statistic for θ.

Solution. The likelihood function of the sample $(X_1, X_2, \ldots, X_n)$ is given by :

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \left[\theta^{\sum_i x_i} (1-\theta)^{n - \sum_i x_i}\right] \times 1$$

$$= g\,[t(\mathbf{x}), \theta]\,.\; h\,(x_1, x_2 \ldots, x_n)$$

where $\quad t(\mathbf{x}) = \sum_{i=1}^{n} x_i \quad$ and $\quad h\,(x_1, x_2, \ldots, x_n) = 1$

Hence by Factorisation Theorem, $T = \sum_{i=1}^{n} X_i$, is sufficient estimator of θ.

Since X_i's are i.i.d. Bernoulli variates with parameter θ,

$$T = \sum_{i=1}^{n} X_i \sim B\,(n, \theta),$$

with p.m.f.

$$P\,(T = k) = \begin{cases} {}^nC_k\, \theta^k (1-\theta)^{n-k}, & k = 0, 1, 2, \ldots, n \\ 0 & , \text{otherwise} \end{cases}$$

$$E_\theta\,[h(T)] = \sum_{k=0}^{n} h(k)\,.\,P(T = k) = \sum_{k=0}^{n} h(k).\,{}^nC_k\, \theta^k (1-\theta)^{n-k}$$

$$= \sum_{k=0}^{n} A(k)\,.\,\theta^k (1-\theta)^{n-k} ; \quad A(k) = h(k).\,{}^nC_k \qquad \ldots(*)$$

$$= A(0)\,(1-\theta)^n + A(1)\,\theta(1-\theta)^{n-1} + \ldots + A(n).\,\theta^n$$

Now

$$E_\theta\,[h(T)] = 0 \text{ for all } \theta \in \Theta = \{\theta : 0 < \theta < 1\}$$

$\Rightarrow \quad A(0)\,(1-\theta)^n + A(1)\,\theta\,(1-\theta)^{n-1} + \ldots + A(n)\,\theta^n = 0, \forall\, \theta$

$\Rightarrow \quad A(0) + A_1\,[\theta/(1-\theta)] + \ldots + A(n)\,[\theta/(1-\theta)]^n = 0\; \forall\, \theta \in [0, 1]$ ...(**)

$\Rightarrow \quad A(0) = A(1) = A(2) = \ldots = A(n) = 0,$

since a polynomial of degree n in x is identically zero (for all x), if all the coefficients are zero.

From (*) and (**), we get

$$h(k) = 0,\ k = 0, 1, 2, \ldots, n$$

$$\Rightarrow \quad h(t) = 0,\ t = 0, 1, 2, \ldots, n$$

Hence T is a complete (sufficient) statistic for θ.

Example 15·25. *Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from $N(\theta, 1)$ population. Examine if $T = t(x) = X_1$ is complete for θ.*

Solution. We have $T = X_1$; $\Theta = \{\theta : -\infty < \theta < \infty\}$

$$\therefore \quad E_\theta[h(T)] = 0$$

$$\Rightarrow \quad \int_{-\infty}^{\infty} h(u)\, e^{-(u-\theta)^2/2}\, du = 0, \text{ for all } \theta \in \Theta$$

$$\Rightarrow \quad \int_{-\infty}^{\infty} \left\{h(u)\, e^{-u^2/2}\right\} e^{\theta u} \,.\, du = 0, \text{ for all } \theta \in \Theta$$

This is a bilateral Laplace transform in θ. Since these are unique :

$$h(u) \,.\, e^{-u^2/2} = 0,\ a.s.$$

$$\Rightarrow \quad h(u) = 0,\ a.s.$$

$$\Rightarrow \quad P[h(T) = 0] = 1,\ \forall\ \theta \in \Theta$$

$$\Rightarrow \quad T = X_1, \text{ is complete statistic for } \theta.$$

Remark. It can be easily seen that $T_1 = \sum_{i=1}^{n} X_i$, is a sufficient estimator of θ and since $T_1 \sim N(n\theta, 1/n)$, by proceeding as in the above problem, we can prove that $T_1 = \sum_{i=1}^{n} X_i$, is a complete sufficient statistic for θ and the family of distributions $\{g_1(t_1, \theta), \theta \in \Theta\}$, is complete.

Example 15·26. *Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(0, \theta)$. Prove that $T = X_1$ is not a complete statistic for θ but $T_1 = X_1^2$ is complete for θ.*

Solution. Here $T = t(x) = X_1$; $\Theta = \{\theta\ ;\ 0 < \theta < \infty\}$

$$E_\theta[h(T)] = 0, \text{ for all } \theta \in \Theta$$

$$\Rightarrow \quad \int_{-\infty}^{\infty} h(u) \exp\left[-u^2/(2\theta)\right] du = 0, \text{ for all } \theta \in \Theta$$

This holds only for all odd functions $h(u)$ of u, for which the integral exists *i.e.*, for all functions *s.t.*

$$
\begin{aligned}
& h(u) && = -h(-u);\ \text{for all } u \\
\Rightarrow\quad & h(u) && \neq 0,\ \text{a.s.} \\
\Rightarrow\quad & T && = X_1 \text{ is not complete statistic for } \theta.
\end{aligned}
$$

Let us now consider the statistic $T_1 = X_1^2$.

$$E_\theta\,[h(T_1)] = 0,\ \text{for all } \theta \in \Theta$$

$$\Rightarrow\quad \int_{-\infty}^{\infty} h(x^2)\exp(-x^2/2\theta)\,dx = 0,\ \text{for all } \theta \in \Theta$$

$$\Rightarrow\quad \int_{-\infty}^{\infty} \frac{h(u)}{\sqrt{u}}\exp(-u/2\theta)\,du = 0,\ \forall\ \theta \in \Theta$$

This being a Laplace transform in $(1/\theta)$, we have

$$
\begin{aligned}
& \frac{h(u)}{\sqrt{u}} && = 0,\ a.s. \\
\Rightarrow\quad & h(u) && = 0,\ a.s. \\
\Rightarrow\quad & T_1 && = X_1^2, \text{ is complete statistic for } \theta.
\end{aligned}
$$

Remark. We can easily see that $T_1 = X_1^2$, is sufficient statistic for θ. Hence $T_1 = X_1^2$ is a complete sufficient statistic for θ.

Example 15·27. *Let $X_1, X_2, \ldots, X_n$ be a random sample from uniform $U[0,\ \theta],\ \theta > 0$ population. Show that $T = \max\limits_{1 \leq i \leq n}(X_i) = X_{(n)}$, is a complete sufficient statistic for θ.*

Solution. $T = X_{(n)}$ has p.d.f.

$$g(t, \theta) = \begin{cases} \dfrac{n\, t^{n-1}}{\theta^n}\ ; & 0 \leq t \leq \theta \\ 0\ , & \text{otherwise} \end{cases}$$

$$E_\theta\,[h(T)] = 0,\ \text{for all } \theta \in \Theta = \{\theta : 0 < \theta < \infty\}$$

$$\Rightarrow\quad \frac{n}{\theta^n}\int_0^{\theta} h(u).\ u^{n-1}\,du = 0,\ \text{for all } \theta \in \Theta$$

Differentiating w.r. to θ, we get from the fundamental theorem of integral calculus :

$$h(\theta)\,.\,\theta^{n-1} = 0,\ \forall\ \theta \in \Theta$$

$$\Rightarrow\quad h(T) = 0,\ a.s.$$

$$\Rightarrow\quad T = \max(X_1, X_2, \ldots, X_n) = X_{(n)}, \text{ is complete for } \theta.$$

We have also proved in Example 15·13, that $T = X_{(n)}$, is sufficient for θ.

Hence $T = X_{(n)}$, is complete sufficient statistic for θ.

15·9. MVU and Blackwellisation. Cramer-Rao inequality (c.f. § 15·7) provides us a technique of finding if the unbiased estimator is also an *MVU* estimator or not. Here, since the regularity conditions are very strict, its applications become quite restrictive. More-over *MVB* estimator is not the same as an *MVU* estimator since the Cramer-Rao lower bound may not always be attained. More-over, if the regularity conditions are violated, then the least attainable variance may be less than the Cramer-Rao bound. [For illustration see Example 15·30]. In this section we shall discuss *how* to *obtain MVU estimator from any unbiased estimator through the use of sufficient statistic.* This technique is called Blackwellisation after D. Blackwell. The result is contained in the following Theorem due to C.R. Rao and D. Blackwell.

Theorem 15·9. (Rao-Blackwell Theorem). *Let X and Y be random variables such that*

$$E(Y) = \mu \text{ and } Var(Y) = \sigma_Y^2 > 0$$

Let $E(Y \mid X = x) = \phi(x)$, *then*

(i) $E[\phi(X)] = \mu$

and (ii) $Var[\phi(X)] \le Var(Y)$

Proof. Let $f_{XY}(x, y)$ be the joint p.d.f. of random variables X and Y, $f_1(.)$ and $f_2(.)$ the marginal p.d.f.'s of X and Y respectively and $h(y \mid x)$ be the conditional p.d.f. of Y for given $X = x$ such that

$$h(y \mid x) = \frac{f(x, y)}{f_1(x)}$$

$$E(Y \mid X = x) = \int_{-\infty}^{\infty} y.\, h(y \mid x) dy$$

$$= \int_{-\infty}^{\infty} y \,.\, \frac{f(x, y)}{f_1(x)} dy$$

$$= \frac{1}{f_1(x)} \int_{-\infty}^{\infty} y f(x, y)\, dy = \phi(x), \text{ (say)} \qquad \ldots(15{\cdot}46)$$

$$\Rightarrow \quad \int_{-\infty}^{\infty} y f(x, y) dy = \phi(x) . f_1(x) \qquad \ldots(15{\cdot}46a)$$

From (15·46) we observe that the conditional distribution of Y given $X = x$ does not depend on the parameter μ. Hence X is sufficient statistic for μ.

Now

$$E[\phi(X)] = E[E(Y \mid X)] = E(Y) = \mu, \qquad \ldots(15{\cdot}47)$$

which establishes part (*i*) of the Theorem.

We have

$$\text{Var}(Y) = E[Y - E(Y)]^2 = E[Y - \mu]^2$$

$$= E[Y - \phi(X) + \phi(X) - \mu]^2$$

$$= E[Y - \phi(X)]^2 + E[\phi(X) - \mu]^2 + 2E[\{Y - \phi(X)\} \{(\phi(X) - \mu\}] \qquad \ldots(15{\cdot}48)$$

The product term gives

$$E[\{Y - \phi(X)\}\ \{\phi(X) - \mu\}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \phi(x))\ (\phi(x) - \mu) f(x, y) dx dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \phi(x))[\phi(x) - \mu] f_1(x)\ h(y \mid x) dx dy$$

$$= \int_{-\infty}^{\infty} [\phi(x) - \mu] \left[\int_{-\infty}^{\infty} [y - \phi(x)] h(y \mid x) dy \right] dx$$

But $\displaystyle\int_{-\infty}^{\infty} [y - \phi(x)] h(y \mid x)\ dy = 0$ $\quad [\because E(Y \mid X = x) = \phi(x)]$

$\therefore \quad E[(Y - \phi(X))(\phi(X) - \mu)] = 0$

Substituting in (15·48), we get

$$\text{Var}\ (Y) = E[Y - \phi(X)]^2 + \text{Var}\ [\phi(X)] \qquad \text{...(15·49)}$$

$$\Rightarrow \quad \text{Var}\ Y \geq \text{Var}\ [\phi(X)] \qquad (\because E[Y - \phi(X)]^2 \geq 0)$$

$$\Rightarrow \quad \text{Var}\ [\phi(X)] \leq \text{Var}\ Y, \qquad \text{...(15·49}a\text{)}$$

which completes the proof of the theorem.

Remarks. 1. From (15·49), it is obvious that the sign of equality holds in (15·49*a*) iff

$$E[Y - \phi(X)]^2 = 0$$

$$\Rightarrow \quad Y - \phi(X) = 0, \text{ almost surely.}$$

i.e., iff $\quad P\{(x, y) : y - \phi(x) = 0\} = 1 \qquad$ ...(15·50)

2. Here we have proved the theorem for continuous r.v.'s. The result can be similarly proved for discrete case, replacing integration by summation.

3. Rao-Blackwell theorem enables us to obtain MVU estimators through sufficient statistic. If a sufficient estimator exists for a parameter, then in our search for MVU estimator we may restrict ourselves to functions of the sufficient statistic. The theorem can be stated slightly different as follows :

Let $U = U(x_1, x_2, \ldots, x_n)$ be an unbiased estimator of parameter $\gamma(\theta)$ and let $T = T(x_1, x_2, \ldots, x_n)$ be sufficient statistic for $\gamma(\theta)$. Consider the function $\phi(T)$ of the sufficient statistic defined as

$$\phi(t) = E(U \mid T = t) \qquad \text{...(15·51)}$$

which is independent of θ (since T is sufficient for $\gamma(\theta)$). Then

$$E\phi(T) = \gamma(\theta)$$

and $\quad Var\ \phi(T) \leq Var\ (U) \qquad$ ...(15·52)

This result implies that starting with an unbiased estimator U, we can improve upon it by defining a function $\phi(T)$ of the sufficient statistic as given in (15·51). This technique of obtaining improved estimators is called Blackwellisation.

If in addition, the sufficient statistic T is also complete, then the estimator $\phi(T)$ discussed above will not only be an improved estimator over U but also the *best (unique)* estimator. We state below the relevant theorem.

Theorem 15·10. *Let T be a complete sufficient statistic for $\gamma(\theta)$, $\theta \in \Theta$. Then $\phi(T)$, the function of T defined in (15·51) is the unique unbiased estimator of $\gamma(\theta)$.*

Combining the results of the two Theorems 15·9 and 15·10, we have the following result.

Corollary. If *T is a complete sufficient statistic for $\gamma(\theta)$ and if we can find some function of T, say $g(T)$, which is unbiased estimator of $\gamma(\theta)$, then $g(T)$ is the MVU estimator of $\gamma(\theta)$.*

Example 15·28. *Let $X_1, X_2, ..., X_n$ be a random sample from $N(\theta, 1)$. Obtain MVUE of θ.*

Solution. it can be easily proved [*c.f.* Example 15·25] that the statistic

$$T = X_1 + X_2 + ... + X_n = \sum_{i=1}^{n} X_i$$

is complete sufficient statistic for θ.

Consider $$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{T}{n} = g(T), \text{ (say)}$$

Since $\bar{X}_n = g(T)$, is unbiased estimator of θ, by corollary to Theorem 15·10, $\bar{X}_n$ is *MVUE* of θ.

Example 15·29. *Let $X_1, X_2, ..., X_n$ be a random sample from $U[0, \theta]$ population. Obtain MVUE for θ.*

Solution. We have seen that in sampling from U $[0, \theta]$ population, the statistic :

$$T = X_{(n)} = \max_{1 \le i \le n} (X_i)$$

is sufficient (Example 15·13) and complete (Example 15·27) for θ. Also

$$E(T) = E[X_{(n)}] = \left(\frac{n}{n+1}\right)\theta \qquad \text{[See Example 15·30]}$$

$$\Rightarrow \qquad E\left[\frac{(n+1)T}{n}\right] = \theta$$

Hence by corollary to Theorem 15·10, $[(n+1)T/n] = [(n+1)X_{(n)}/n]$ is an *MVU* estimator of θ.

Example 15·30. *Given :*

$$f(x, \theta) = \frac{1}{\theta}, \; 0 < x < \theta, \; \theta > 0 \qquad ...(*)$$
$$= 0, \text{ elsewhere,}$$

compute the reciprocal of

$$n\, E\left\{\left[\frac{\partial \log f(x, \theta)}{\partial \theta}\right]^2\right\}$$

and compare this with the variance of $(n+1)Y_n/n$, where Y_n is the largest item of a random sample of size n from this distribution. Comment on the result.

Solution. $\log f(x, \theta) = -\log \theta \Rightarrow \dfrac{\partial}{\partial\theta}\log f = -\dfrac{1}{\theta}$,

$$\Rightarrow \qquad n\,E\left(\frac{\partial}{\partial\theta}\log f\right)^2 = n\,E\left(\frac{1}{\theta^2}\right) = \frac{n}{\theta^2}$$

Hence reciprocal of $n\,E\left\{\left[\dfrac{\partial}{\partial\theta}\log f(x, \theta)\right]^2\right\} = \dfrac{\theta^2}{n}$ ...(**)

For the rectangular population (*), the *p.d.f.* of nth order statistic (the largest sample observation), Y_n is

$$g(y) = n \cdot [F(y, \theta)]^{n-1} f(y, \theta)$$

where $\quad F(x, \theta) = P(X \le x) = \displaystyle\int_0^x f(u)du = \int_0^x \frac{1}{\theta} = \frac{x}{\theta}$

$$g(y) = n\left(\frac{y}{\theta}\right)^{n-1}\frac{1}{\theta} = \frac{n}{\theta^n}\cdot y^{n-1};\ 0 \le y < \theta$$

$$E(Y_n^r) = \int_0^\theta y^r \cdot g(y)dy = \frac{n}{\theta^n}\int_0^\theta y^{r+n-1}\,dy = \frac{n\theta^r}{n+r}$$

Taking $r = 1$ and 2, we get

$$E(Y_n) = \frac{n\theta}{n+1};\ E(Y_n^2) = \frac{n\theta^2}{n+2} \qquad ...(***)$$

Now $\quad E\left[\dfrac{n+1}{n}\cdot Y_n\right] = \dfrac{n+1}{n}E(Y_n) = \theta \qquad$ [Using ***]

$\Rightarrow \qquad (n+1)Y_n/n$ is an unbiased estimator of θ.

$$\text{Var}\left[\frac{n+1}{n}Y_n\right] = \left(\frac{n+1}{n}\right)^2 \cdot \text{Var}(Y_n)$$

$$= \left(\frac{n+1}{n}\right)^2 [EY_n^2 - (EY_n)^2]$$

$$= \left(\frac{n+1}{n}\right)^2\left[\frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2}\right] \qquad \text{[Using (***)]}$$

$$= \theta^2\left[\frac{(n+1)^2}{n(n+2)} - 1\right] = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n}$$

$$\Rightarrow \qquad \text{Var}\left[\frac{n+1}{n}\cdot Y_n\right] \le 1\Bigg/\left[n\,E\left(\frac{\partial}{\partial\theta}\log f\right)^2\right]$$

Hence $\qquad (n+1)Y_n/n$ is an *MVUE*.

Remark. This example illustrates that if the regularity conditions underlying Cramer-Rao inequality are violated, then the least attainable variance may be less than the Cramer-Rao lower bound.

EXERCISE 15(a)

1. What do you understand by Point Estimation ? Define the following terms and give one example for each :

(*i*) Consistent Statistic
(*ii*) Unbiased Statistic
(*iii*) Sufficient Statistic
(*iv*) Efficiency. [*Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1982*]

2. What do you understand by Point Estimation ? When would you say that estimate of a parameter is good ? In particular, discuss the requirements of consistency and unbiasedness of an estimate. Give an example to show that a consistent estimate need not be unbiased.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986*]

3. Discuss the terms (*i*) estimate, (*ii*) consistent estimate, (*iii*) unbiased estimate, of a parameter and show that sample mean is both consistent and unbiased estimate of the population mean.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

4. (*a*) If $s_1^2, s_2^2, \ldots, s_r^2$ are r sample variances based on random samples of sizes $n_1, n_2, \ldots, n_r$ respectively, and if T is some statistic given by

$$T = \frac{n_1 s_1^2 + n_2 s_2^2 + \ldots + n_r s_r^2}{a},$$

for estimating σ^2 as an unbiased estimator, find the value, of a, supposing population is very large and for every sample

$$s^2 = \frac{1}{n} \Sigma(x_i - \bar{x})^2$$

Ans. $a = (n_1 + n_2 + \ldots + n_r) - r$.

(*b*) If $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_r$ are the sample means based on samples of sizes $n_1, n_2, n_3, \ldots, n_r$ respectively, an unbiased estimator

$$t = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \ldots + n_r \bar{X}_r}{k}$$

has been defined to estimate μ. Find the value of k.

Ans. $k = n_1 + n_2 + \ldots + n_r$.

5. (*a*) For the geometric distribution,

$$f(x, \theta) = \theta\, (1 - \theta)^{x-1},\ (x = 1, 2, \ldots), 0 < \theta < 1,$$

Obtain an unbiased estimator of $1/\theta$. [**Ans.** $E(\bar{X}) = 1/\theta$.]

(*b*) The random variable X takes the values 1 and 0 with respective probabilities θ and $1 - \theta$. Independent observations $X_1, X_2, \ldots, X_n$ on X are available. Write $\xi = X_1 + X_2 + \ldots + X_n$.

Show that $\xi\,(n - \xi)/n(n - 1)$ is an unbiased estimate of $\theta(1 - \theta)$.

6. Show that if T is an unbiased estimator of a parameter θ, then $\lambda_1 T + \lambda_2$ is an unbiased estimator of $\lambda_1 \theta + \lambda_2$, where λ_1 and λ_2 are known constants, but T^2 is a biased estimator of θ^2.

7. For the following cases determine if the given estimator is unbiased for the parametric function. When it is biased, derive an unbiased estimator from it. $\bar{x}$ is the sample mean.

(a) $x_1, \ldots, x_n$ is a random sample from a distribution with variance σ^2. The estimator $n^{-1}[(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2]$
is used to estimate σ^2.

(b) $x_1, \ldots, x_n$ is an independent sample from an exponential distribution with mean θ. The estimator $\left(1 - \dfrac{1}{n\bar{X}}\right)^{n-1}$ is used to estimate $\exp\left(-\dfrac{1}{\theta}\right)$ when $n\bar{X} > 1$ and zero is used when $n\bar{X} < 1$.

(c) r successes are observed in n Bernoulli trials with success probability p. $(r/n)^2$ is used to estimate p^2.

8. $$f(x\,;\mu,\sigma) = \frac{1}{\sigma}\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right];\quad \mu \leq x < \infty,\ -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty$$

Obtain

(i) an unbiased estimate of μ when σ is known,

(ii) an unbiased estimate of σ when μ is known,

(ii) two unbiased estimators of σ^2 when μ is known.

Hence obtain an infinity of unbiased estimators of σ^2 in this case.

[**Hint.** The exponential distribution has mean $\mu + \sigma$ and variance σ^2]

9. Suppose X and Y are independent random variables with the same unknown means μ. Both X and Y have variance as 36. Let $T = aX + bY$ be an estimator of μ.

(i) Show that T is an unbiased estimator of μ if $a + b = 1$.

(ii) If $a = \frac{1}{3}$ and $b = \frac{2}{3}$, what is the variance of T ?

(iii) If $a = \frac{1}{2}$ and $b = \frac{1}{2}$, what is the variance of T ?

(iv) What choice of a and b minimizes the variance of T subject to the requirement that T is an unbiased estimate of μ ?

10. (a) Examine the unbiasedness of the following estimates :

(i) $s_1^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$

(ii) $s_2^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - \mu)^2$ (where μ is known),

for σ^2, the population variance.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1982*]

Ans. $E(s_1^2) = \left(\dfrac{n-1}{n}\right)\sigma^2 \neq \sigma^2$, (ii) $E(s_2^2) = \sigma^2$.

(b) Let $x_1, x_2, \ldots, x_n$ be a random sample of size n drawn from a population with mean μ and variance σ^2. Obtain an unbiased estimator for μ^2.

Hint. $E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$; $\text{Var}(\bar{X}) = \sigma^2/n$

$$E(\bar{X}^2) = \text{Var}(\bar{X}) + [E(\bar{X})]^2 = \mu^2 + (\sigma^2/n)$$

Ans. $\bar{X}^2 - (\sigma^2/n)$, if σ^2 is known ;

and $\bar{X}^2 - (S^2/n) = \bar{X}^2 - \dfrac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$, if σ^2 is unknown.

11. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from $N(\mu, \sigma^2)$, where μ is known and if

$$T = \frac{1}{n}\sum_{i=1}^{n} |X_i - \mu|$$

examine if T is unbiased for σ. If not, obtain an unbiased estimator of σ.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1987*]

Hint. $E(T) = \frac{1}{n}\sum_{i=1}^{n} E|X_i - \mu| = \sqrt{(2/\pi)}\,.\,\sigma$,

since, for $N(\mu, \sigma^2)$, Mean Deviation about mean $= \sqrt{(2/\pi)}\ \ \sigma$

Ans. No; $\sqrt{(\pi/2)}\ T$.

12. If $x_1, x_2, \ldots, x_n$ is a random sample from the population

$$f(x, \theta) = (\theta + 1)\, x^{\theta}\ ;\ 0 < x < 1\ ;\ \theta > -1 \qquad \ldots(*)$$

show that $\left[\dfrac{-(n-1)}{\Sigma \log x_i} - 1\right]$ is unbiased estimator of θ.

Hint. In sampling from (*), $U = -\log X$ has an exponential distribution with parameter $(\theta + 1)$

$$\Rightarrow \qquad U_i = -\log X_i \overset{\text{i.i.d.}}{\sim} \gamma(\theta + 1, 1)\ ;\ i = 1, 2, \ldots, n$$

$$\Rightarrow \qquad Y = -\sum_{i=1}^{n} \log X_i \sim \gamma(\theta + 1, n);\ E\,[1/Y] = \frac{\theta + 1}{n - 1}$$

13. Suppose X has a truncated Poisson distribution with p.m.f.

$$f(x, \theta) = \begin{cases} \dfrac{\exp(-\theta)\,.\,\theta^x}{[1 - \exp(-\theta)]\, x\,!}, & x = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Show that the only unbiased estimator of $[1 - \exp(-\theta)]$ based on X is the statistic T, defined as :

$$T(x) = \begin{cases} 0, & \text{when } x \text{ is odd} \\ 2, & \text{when } x \text{ is even} \end{cases}$$

Note. This is an Example of *absurd unbiased estimator*.

14. Consider a random sample X_1, X_2, X_3 of size 3 from uniform p.d.f.

$$f(x, \theta) = \begin{cases} 1/\theta, & 0 < x < \theta \\ 0, & \text{otherwise} \end{cases}$$

Show that each of the statistics $4X_{(1)}$, $2X_{(2)}$ and $\frac{1}{3}X_{(3)}$, where $X_{(i)}$ is the ith order statistic is an unbiased estimator for θ. Find the variance and hence the efficiency of each.

15. Obtain an unbiased estimator for (*i*) θ, and (*ii*) θ^2, in case of binomial probability distribution :

$$f(x, \theta) = {}^nC_x\, \theta^x (1-\theta)^{n-x} ; x = 0, 1, 2, \dots, n ; 0 < \theta < 1.$$

Hint. $E\left(\dfrac{x}{n}\right) = \theta$; $E\left[\dfrac{x(x-1)}{n(n-1)}\right] = \theta^2$.

If we write $T = x/n$, the observed proportion of successes then

$$E(T) = \theta ; E(T^2) = \frac{\theta^2}{n} + \left(\frac{n-1}{n}\right) . \theta^2 \neq \theta^2.$$

This illustrates that we may have :

t_n unbiased for θ but t_n^2 not unbiased for θ^2.

16. Define 'efficiency of an estimator'.

X is a uniform random variable with range $[0, \theta]$. $x_1, x_2, \dots, x_n$ are independent observations on X. Define

$$\hat{\theta}_1 = \frac{2}{n}(x_1 + x_2 \dots + x_n) ; \hat{\theta}_2 = \left[\frac{(n+1)}{n}\right] \max(x_1, x_2, \dots, x_n).$$

Show that $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased for θ. Evaluate their relative efficiency.

17. (*a*) The observations $x_1, x_2, \dots, x_n$ represent a random sample from a uniform distribution over the interval $(0, \theta)$, where θ is an unknown parameter.

The statistics $\bar{X}$, m and M are the mean, the smallest value and the largest value respectively for the sample. Find values for k so that, kt is an unbiased estimator for θ where

(*a*) $t = \bar{X}$

(*b*) $t = M$,

(*c*) $t = M - m$,

Of the three unbiased estimators which is the best ? Give your reasons.

(*b*) Let $X_1, X_2, \dots, X_n$ $(n > 2)$ be a random sample of size n from the distribution having density function :

$$f(x ; \theta) = \theta x^{\theta - 1}, 0 < x < 1, \ \theta > 0$$

If $Z = -\sum_{i=1}^{n} \log X_i$, show that $\dfrac{n-1}{Z}$ is an unbiased estimator for θ and its efficiency is $(n-2)/n$.

Hint. See hint to Question 12.

18. (*a*) Suppose $X_1, X_2, \dots, X_n$ are sample values independently drawn from population with mean m and variance σ^2. Consider the estimates :—

$$Y_n = \frac{X_1 + X_2 + \ldots + X_n}{n + 1}, \quad Z_n = \frac{X_1 + 2X_2 + 3X_3 + \ldots + nX_n}{n^2}.$$

Discuss whether they are unbiased, consistent for m. What is the efficiency of Y_n over Z_n ?

(*b*) Let X_1, X_2, X_3 and X_4 be independent random variables such that $E(X_i) = \mu$ and Var $(X_i) = \sigma^2$ for $i = 1, 2, 3, 4$.

If $$Y = \frac{X_1 + X_2 + X_3 + X_4}{4}, \quad Z = \frac{X_1 + X_2 + X_3 + X_4}{5}$$

and $$T = \frac{X_1 + 2X_2 + X_3 - X_4}{4},$$

examine whether Y, Z and T are unbiased estimators of μ ? What is the efficiency of Y relative to Z ?

(*c*) Let x_1, x_2, x_3, x_4, be a random sample from a $N(\mu, \sigma^2)$ population. Find the efficiency of $T = \frac{1}{7}(x_1 + 3x_2 + 2x_3 + x_4)$ relative to $\overline{X} = \frac{1}{4}\sum_1^4 x_i$. Which is relatively more efficient ? Why ?

19. A simple random sample of size 2 is drawn from a population containing 3 units, without replacement. Let y_1, y_2, y_3 be the value of a characteristic measured on the three units and let T_{ij} be the estimator of the population mean $\overline{Y}$ for the sample that has units i and j ; $i, j = 1, 2, 3, i \neq j$.

If $T_{12} = (y_1 + y_2)/2$, $T_{13} = (y_1/2) + (2y_3/3)$, $T_{23} = (y_2/2) + (y_3/3)$, show that T_{ij} is unbiased for $\overline{Y}$. Find the variance of T_{ij} and hence show that the variance of T_{ij} is smaller than that of the sample mean estimator if $y_3(3y_2 - 3y_1 - y_3) = 0$. **[Indian Forest Service, 1991]**

20. Let x, the earnings of a commercial bank, be a random variable with mean μ and variance σ^2. A random sample of earnings of n banks is denoted by $x_1, x_2, \ldots, x_n$. However, because of the disclosure laws, individual bank earnings are not disclosed and only the following average values are made available to the researcher :

$$a_1 = \frac{x_1 + x_2}{2}, a_2 = \frac{x_3 + x_4}{2}, \ldots, a_m = \frac{x_{n-1} + x_n}{2},$$

where n is an even number and $m = n/2$.

(*i*) Devise the best linear unbiased estimator of μ, given the available information. What is the variance of the proposed estimator ?

(*ii*) Devise an unbiased estimator of σ^2. **[Delhi Univ. M.A. (Eco.), 1990]**

21. (*a*) Define a consistent estimator.

Let T_n be an estimator of θ with variance σ_n^2 and $E(T_n) = \theta_n$. Prove that if $\theta_n \to \theta$ and $\sigma_n^2 \to 0$, as $n \to \infty$ then T_n is a consistent estimator of θ.

Hence obtain consistent estimators for :

(*i*) Mean of the normal distribution.

(*ii*) Variance of the normal distribution when mean is known.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

(*b*) Give an example of an estimator :

(*i*) which is consistent but not unbiased,

(*ii*) which is unbiased but not consistent.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

22. (*a*) State and prove a sufficient condition for the consistency of an estimator. Define the invariance property of a consistent estimator and establish it. [*Delhi Univ. B.Sc. (Stat. Hons.), 1985*]

(*b*) Given a random sample $X_1, X_2, \ldots, X_n$ from a normal (μ, σ^2) distribution, examine unbiasedness and consistency of

$$(i)\ \bar{X} \text{ for } \mu, \quad (ii)\ \frac{1}{n} \sum (X_i - \bar{X})^2 \text{ for } \sigma^2.$$

23. (*a*) When would you say that estimate of a parameter is good ? In particular, discuss the requirements of consistency and unbiasedness of an estimate. Give an example to show that a consistent estimate need not be unbiased.

Show that an unbiased estimator whose variance tends to zero as the sample size increases to infinity is consistent.

(*b*) Define unbiasedness and consistency of estimators. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Propose three estimators of μ based on this random sample such that the first is unbiased but not consistent, the second is consistent but not unbiased and the third is both unbiased and consistent. [*Punjab Univ. M.A. (Eco.), 1990*]

24. (*a*) Define an unbiased and consistent estimate of a parameter in a population distribution.

Prove that for a sample of size n from a normal $(m, 1)$ population, the arithmetic mean is an unbiased estimate of m and by Chebyshev's inequality or otherwise, show that the estimate is consistent too.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1991*]

(*b*) If $X_1, X_2, \ldots, X_n$ is a random sample obtained from the density function :

$$\begin{aligned} f(x, \theta) &= 1, \quad \theta < x < \theta + 1 \\ &= 0, \quad \text{elsewhere} \end{aligned}$$

show that the sample mean $\bar{X}$ is an unbiased and consistent estimator of $\theta + \frac{1}{2}$.

25. (*a*) Define a consistent estimator. Let $T_{1,n}$ and $T_{2,n}$ be consistent estimators of $g_1(\theta)$ and $g_2(\theta)$ respectively. Prove that $aT_{1,n} + b\,T_{2,n}$ is a consistent estimator of $ag_1(\theta) + bg_2(\theta)$, where a and b are constants independent of θ.

(*b*) Define consistent estimator. If the estimator t_n based on a random sample of size n is such that

$$E(t_n) \to \theta$$

and

$$V(t_n) \to 0,$$

as $n \to \infty$, then prove that t_n is a consistent estimator for θ. Hence prove that sample mean is always a consistent estimate for population mean.

[*Delhi Univ. M.Sc. (Maths), 1990*]

(*c*) If t_n is a biased estimate of paramater θ based on a random sample of size n, and $E(t_n) = \theta + b_n$ and if $b_n \to 0$ and $V(t_n) \to 0$ as $n \to \infty$, show that t_n is consistent estimator of θ.

(*d*) Define a consistent estimator of parameter θ. If T is a consistent estimator of θ and if ϕ is any continuous function of its argument, show that $\phi(T)$ is a consistent estimator of $\phi(\theta)$.

26. (*a*) Show that $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$, are joint consistent estimators for μ and σ^2 respectively, if $x_1, x_2, \ldots, x_n$ is a random sample from a normal population $N(\mu, \sigma^2)$.

Also find the efficiency of $ns^2/(n-1)$.

(*b*) Show that if t is a consistent estimator of a parameter θ, then e^t is a consistent estimator of e^θ.

(*c*) Prove that in case of Binomial distribution with parameter θ, t_n defined as r/n is a consistent unbiased estimator for θ, but t_n defined as $(r/n)^2$ is consistent but not unbiased estimator for θ^2.

27. Show that in sampling from Cauchy distribution

$$f(x, \theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \quad -\infty < x < \infty, \quad \theta > 0;$$

(*i*) Sample mean $\bar{X}$ is not a consistent estimator of θ.

(*ii*) Sample median is a consistent estimator of θ and its asymptotic efficiency is $8/\pi^2$.

28. (*a*) If T_1 and T_2 are consistent estimators of $\gamma(\theta)$, show that $a_1 T_1 + a_2 T_2$, such that $a_1 + a_2 = 1$, is also consistent for $\gamma(\theta)$.

(*b*) For a Poisson distribution with parameter θ, show that $1/\bar{X}$ is consistent estimator of $1/\theta$, where $\bar{X}$ is the mean of a random sample from the given population.

Hint. Prove that $\bar{X}$ is a consistent estimator of θ and then use Invariance Property (Theorem 15·1).

29. Define *MVU* estimator. If T_1 and T_2 are two unbiased estimators of a parameter θ, with variances σ_1^2 and σ_2^2 and correlation coefficient ρ, then obtain the best unbiased linear combination of T_1 and T_2. Also obtain its variance. [*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

30. (*a*) Let T_1 and T_2 be two unbiased estimators of $\gamma(\theta)$ having the same variance. Show that their correlation coefficient ρ_θ cannot be smaller than $(2e_\theta - 1)$, where e_θ is the efficiency of each estimator.

Further show that if T_1 is *MVU* estimator and T_2 is any unbiased estimator with efficiency e, then

$$V(T_1 - T_2) = \left(\frac{1}{e} - 1\right) V(T_1)$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

(*b*) If T_1 is a *MVU* for θ and T_2 is any other unbiased estimator of θ with efficiency e_θ then prove that the correlation between T_1 and T_2 is $\sqrt{e_\theta}$.

[Delhi Univ. B.A. (Stat. Hons.), 1987]

31. (*a*) Define *MVU* estimator. Show that an *MVU* estimator is unique.

[Delhi Univ. B.Sc. (Stat. Hons.), 1985]

(*b*) If T_1 and T_2 are two unbiased statistics having the same variance and ρ is the correlation between them then show that $\rho \geq 2e - 1$, where e is the ratio of the variance of the best estimator to the common variance of T_1 and T_2.

[Delhi Univ. B.Sc. (Stat. Hons.), 1992]

32. (*a*) Let T be an *MVU* estimate for $\gamma(\theta)$ and T_1, T_2 be two other unbiased estimators of $\gamma(\theta)$ with efficiencies e_1 and e_2 respectively.

If ρ_θ is the correlation coefficient between T_1 and T_2, then

$$(e_1 e_2)^{1/2} - \{(1 - e_1)(1 - e_2)\}^{1/2} \leq \rho_\theta \leq (e_1 e_2)^{1/2} + \{(1 - e_1)(1 - e_2)\}^{1/2}.$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1988, 1986]

(*b*) Let t_1 and t_2 be two unbiased estimates of θ with variances σ_1^2 and σ_2^2, (both known) and correlation ρ (known). Consider the estimate

$$\hat{\theta} = \alpha t_1 + (1 - \alpha)\, t_2.$$

Show that $\hat{\theta}$ is unbiased. Find α such that $\hat{\theta}$ has minimum variance.

[Delhi Univ. M.A. (Eco.), 1986]

33. Suppose X and Y are independent unbiased estimates of μ. It is known that the variance of X is 12 and the variance of Y is 4. It is desired to combine two estimators in order to obtain a more efficient estimator: Let $T = aX + bY$, be the new estimator.

(*i*) In order that T be an unbiased estimator of μ, what conditions must be imposed on a and b ?

(*ii*) Find the values of a and b that minimize the variance of T subject to the condition that T be an unbiased estimator.

34. (*a*) What is an efficient estimator ?

If T_1, T_2 are both efficient estimators with variance v and if $T = \frac{1}{2}(T_1 + T_2)$, show that variance of T is $(v/2)(1 + \rho)$, where ρ is the coefficient of correlation between T_1 and T_2. Deduce that $\rho = 1$ and that T is also efficient.

(*b*) If T and T' be two consistent estimators of which T is the most efficient, prove that the correlation coefficient between them is

$\sqrt{\dfrac{V(T)}{V(T')}}$, where $V(T)$ and $V(T')$ are the variance of T and T' respectively.

Show also that the correlation coefficient between two most efficient estimators is unity.

[Allahabad Univ. M.A. (Eco.), 1993]

35. Define a sufficient statistic. Explain the method of finding sufficient estimator. If $(x_1, x_2, \ldots, x_n)$ is a random sample from a distribution :

$$f(x, p) = p^x (1-p)^{1-x} \,;\, x = 0, 1 \text{ and } 0 \le p \le 1,$$

find the sufficient estimator of p. [*Madras Univ. B.Sc., 1988*]

36. State the factorisation theorem on sufficiency. Obtain a sufficient statistic for the parameter θ in the following distribution :

$$f(x : \theta) = \frac{1}{\theta}, \quad 0 < x < \theta.$$

(*b*) Define a sufficient statistic.

If $x_1, x_2, \ldots, x_n$ is a random sample from a distribution :

$$f(x, \theta) = \theta^x (1-\theta)^x \,;\, x = 0, 1, 0 < \theta < 1$$
$$= 0, \text{ elsewhere.}$$

Show that $Y_1 = x_1 + x_2 + \ldots + x_n$, is a sufficient statistic for θ.

[*Madras Univ. B.Sc., 1987*]

(*c*) Let $x_1, x_2, \ldots, x_n$ denote a random sample from a population with p.d.f

$$f(x, \theta) = \theta x^{\theta - 1}, 0 < x < 1.$$

Show that $Y = x_1 x_2 \ldots x_n$, is a sufficient statistic for θ.

37. (*a*) Let X be a random sample of size one from a normal distribution $N(0, \sigma^2)$.

(*i*) Is X a sufficient statistic for σ^2 ?

(*ii*) Is $|X|$ a sufficient statistic for σ^2 ?

(*iii*) Is X^2 a sufficient statistic for σ^2 ? (*Gujarat Univ. B.Sc., 1992*)

(*b*) Examine which of the following distributions admit sufficient estimators for their parameters :

(*i*) $f(x, \theta) = \theta x^{\theta - 1}, 0 \le x \le 1$

(*ii*) $f(x\, y, \rho) = \dfrac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{-\dfrac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\}$

38. (*a*) Show that if a sufficient estimator exists, it is also the maximum likelihood estimator. Is the converse true ? Explain.

(*b*) Do the following distributions admit of sufficient estimators ?

(*i*) $f(x, \theta) = \dfrac{1}{\theta} \,;\, k\theta \le x \le (k+1)\theta$, where k is an integer.

(*ii*) $f(x, \theta) = \dfrac{1+\theta}{(x+\theta)^2}, 1 \le x < \infty$

39. (*a*) Prove that if an unbiased estimator and a sufficient statistic exist for $\psi(\theta)$ and the density function $f(x, \theta)$ satisfies certain regularity conditions (to be stated by you), then the best unbiased estimate of $\psi(\theta)$ is an explicit function of the sufficient statistic.

Examine if the following distribution admits a sufficient statistic for the parameter θ.

$$f(x, \theta) = (1+\theta)\, x^\theta \,;\, 0 \le x \le 1, \theta > 0$$

(*b*) Discuss if a sufficient statistic exists for the parameter θ, in sampling from double exponential distribution with p.d.f.

$$f(x, \theta) = \frac{1}{2} \exp(-|x - \theta|), -\infty < x < \infty.$$

Hint. Proceed as in Example 15·17.

Ans. No sufficient estimator for θ exists.

(*c*) Obtain jointly sufficient estimators for α and β in a random sample $X_1, X_2, \ldots, X_n$ from the uniform population with p.d.f.

$$f(x, \alpha, \beta) = \frac{1}{\beta - \alpha}, \quad \alpha \le x \le \beta$$
$$= 0 \quad , \text{otherwise}$$

Ans. $T_1 = X_{(1)}$ and $T_2 = X_{(n)}$, are jointly sufficient for α and β respectively.

40. (*a*) Show that a necessary and sufficient condition for a statistic T to be sufficient for θ is that the probability function $f_\theta(x)$ should belong to an exponentially family.

(*b*) Let $x_1, x_2, \ldots x_n$ be a random sample from a distribution with p.d.f. $f(x : \theta) = e^{-(x-\theta)}, x \ge \theta, -\infty < \theta < \infty$. Obtain a sufficient statistic for θ.

[Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1985]

41. Define a sufficient statistic. State and prove the Factorisation theorem on sufficiency. *[Delhi Univ. B.Sc. (Stat. Hons.), 1986]*

42. (*a*) Let (X_1, X_2, X_3) be a random sample from the probability mass function : $P(X = x) = \theta^x (1 - \theta)^{1-x}, (x = 0, 1; 0 < \theta < 1)$.

If $t = X_1 + X_2 + X_3$, show that the conditional distribution of the random sample given $t = r$, does not depend on θ. Interpret this result in the light of sufficiency-concept.

(*b*) Let (X_1, X_2) be a random sample from a Poisson distribution with parameter θ. Prove that $t = X_1 + 2X_2$ is not sufficient for θ.

(*c*) Let (X_1, X_2) be a random sample from $N(\theta, 1)$. If $T = X_1 + X_2$ and $U = X_2 - X_1$, show that the conditional distribution of U given $T = t$, does not depend on θ. Interpret this result in the light of sufficiency-concept.

(*d*) For a random sample X_i $(i = 1, 2, \ldots, n)$, from an exponential distribution with p.d.f.

$$f(x, \theta) = \frac{1}{\theta} \exp\left[-\frac{x}{\theta}\right], \quad x > 0, \quad \theta > 0,$$

obtain an unbiased and sufficient estimator for θ.

[Delhi Univ B.Sc. (Stat. Hons.) 1983, 1988]

43. Prove that under certain regularity conditions to be stated by you, the variance of an unbiased estimator T for $\gamma(\theta)$, satisfies the inequality

$$\text{Var}_\theta(T) \ge \frac{[\gamma'(\theta)]^2}{E_\theta\left[\frac{\partial \log f_\theta(X_1, X_2, \ldots, X_n)}{\partial \theta}\right]^2}.$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986]

44. (*a*) If T is an unbiased estimator of a parameter θ, based on a random sample of size n, prove that

Var $(T) \geq 1/[nI(\theta)]$, where $I(\theta)$ is the information function.

(*b*) Show that under certain regularity conditions, an unbiased estimate T of a parametric function $\psi(\theta)$ attains a Cramer-Rao bound for the variance of unbiased estimator of $\psi(\theta)$, if and only if T satisfies the relation

$$\frac{\partial \log L}{\partial \theta} = \frac{n\,I(\theta)}{\psi'(\theta)} \{T - \psi(\theta)\}$$

where L is the likelihood function of a sample of n observations and

$$nI(\theta) = E\left(\frac{\partial \log L}{\partial \theta}\right)^2$$

What is the variance of T in such a case ? Show that an estimator T satisfying the above relation is unique when it exists. Further a parametric function admitting such an estimator T is unique except for an additive and multiplicative constant. *(Meerut Univ. B.Sc., 1992)*

45. (*a*) State and Prove Cramer-Rao Inequality.

(*b*) Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with p.d.f.

$$f(x, \theta) = \theta \cdot e^{-\theta x} \; ; \; x > 0, \theta > 0.$$

Find Cramer-Rao lower bound for the variance of the unbiased estimator of θ. *[Delhi Univ. B.Sc. (Stat. Hons.), 1987]*

46. $f(x, \theta)$ is a probability density function and $(x_1, x_2, \ldots, x_n)$ is a random sample from it. Prove that if an unbiased minimum variance bound (*MVB*) estimator T exists, it must be of the form $T = \theta + \lambda \sum_i \frac{\partial}{\partial \theta} \log f(x_i, \theta)$, in which λ does not depend on sample values.

Show that the variance of T is λ and is given by

$$\frac{1}{\text{Var } T} = n\, E\left\{\frac{\partial^2}{\partial \theta^2} \log f(x_i, \theta)\right\}$$

Write a note on the connection between *MVB* estimators and sufficiency, giving example.

47. (*a*) Define Minimum Variance unbiased estimator and Minimum Variance Bound unbiased estimator and explain clearly the difference between them. Prove that minimum variance unbiased estimator is essentially unique.

(*b*) Verify that there exists an M.V.B. estimator for the parameter θ of the distribution :

$$f(x, \theta) = \frac{e^{-\theta} \cdot \theta^x}{x\,!} \; ; \; x = 0, 1, 2, \ldots$$

and hence obtain the value of M.V.B. *(Marathwada Univ. M.Sc., 1993)*

(*c*) Show that there exists a parameter function $\psi(\theta)$ in the case of the geometric distribution :

$$f(x, \theta) = (1 - \theta)\, \theta^x \; ; \; x = 0, 1, 2, \ldots ; \; 0 < \theta < 1$$

such that there exists an M.V.B. unbiased estimator T of $\psi(\theta)$.

Obtain $\psi(\theta)$, T and $V(T)$. *[Agra Univ. M.Sc., 1988]*

48. (*a*) Define minimum variance unbiased estimator (*MVUE*). How is Cramer-Rao inequality useful in obtaining such an estimator ? Derive this inequality.

(*b*) Obtain minimum variance unbiased estimator of θ from a sample of n independent observations $x_1, x_2, \ldots, x_n$ drawn from the binomial $B\ (N, \theta)$ population having probability function :

$$f(x\ ;\theta) = {}^N C_x\, \theta^x (1-\theta)^{N-x},\ x = 0, 1, 2, \ldots, N.$$

Also obtain variance of this estimator of θ.

49. (*a*) If $b(\theta)$ is the bias in the estimator T of θ, then show that (under conditions to be stated by you),

$$E(T-\theta)^2 \geq \frac{\{1+b'(\theta)\}^2}{I(\theta)} + \{b(\theta)\}^2,$$

where $I(\theta)$ is the information on θ supplied by a sample of n observations.

(*b*) Prove the following result :

$$\int_{-\infty}^{\infty} (x-\theta)^2 \, .\, g(x, \theta)\, dx \int_{-\infty}^{\infty} \left(\frac{\partial \log g}{\partial\theta}\right)^2 g(x, \theta)\, dx \geq \left(\frac{d\psi}{d\theta}\right)^2$$

where $g(x, \theta)$ is the frequency function in x having the first moment $\psi(\theta)$ and finite second moment. Discuss when the equality sign holds.

50. For the gamma distribution

$$f(x, \theta) = \frac{1}{\theta^p \Gamma p} x^{p-1} \exp(-x/\theta)\ ;\ 0 \leq x < \infty, \theta > 0,\ p \text{ (known)},$$

find the expectation of X^2. Use it to obtain an unbiased estimator T of θ^2. Find $V(T)$.

Evaluate Fisher's information function $I(\theta)$ about θ^2 and verify the truth of the inequality :

$$V(T) > [n\, I(\theta^2)]^{-1}$$

51. State and prove Rao-Blackwell theorem and explain its significance in the theory of point estimation.

Let $x_1, x_2, \ldots, x_n$ be a random sample from Poisson distribution with parameter λ. Obtain Cramer-Rao lower bound to the variance of an unbiased estimator for λ. Hence find the M.V.U.E. for λ.

[*Delhi Univ. M.Sc., (Maths.), 1990*]

52. State and prove Rao-Blackwell theorem and explain its significance in point estimation.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a rectangular distribution with p.d.f.

$$f(x, \theta) = 1/\theta,\ \ 0 \leq x \leq \theta.$$

Find MVU estimators of θ and $3\theta + 5$.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1987*]

53. Define completeness of a statistic T. Let $X_1, X_2, \ldots, X_n$ be a random sample from uniform population $U[0, \theta]$. Obtain sufficient statistic for θ. Show that it is complete. Hence obtain MVU estimator for θ.

[*Delhi Univ B.Sc. (Stat. Hons.), 1988*]

54. Define a complete sufficient statistic.

If T is a complete sufficient statistic for $\gamma(\theta)$, and $E[\phi(T)] = \gamma(\theta)$, then show that $\phi(T)$ is the unique $MVUE$ of $\gamma(\theta)$.

Use this property and obtain MVU estimator of θ based on a random sample $X_1, X_2, ..., X_n$ from the distribution with p.m.f.

$$f(x, \theta) = \begin{cases} \theta^x (1-\theta)^{1-x}, & x = 0, 1 \\ 0 & , \text{elsewhere} \end{cases}$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

55. Show that the family $\{f(x, \theta), \theta \in (0, 1)\}$ with

$$f(x, \theta) = {}^2C_x \theta^x (1-\theta)^{2-x}, \quad x = 0, 1, 2,$$

is complete. [*Delhi Univ. B.Sc. (Stat. Hons.), 1993*]

56. Let $X_1, X_2, ..., X_n$ be a random sample from

$$f_\theta(x) = \frac{1}{\theta}, 0 < x < \theta \text{ for all } \theta \in \Theta$$

Show that $X_{(n)} = \max(X_1, X_2, ..., X_n)$ is sufficient for θ and $\dfrac{(n+1)}{n} X_{(n)}$ is an unbiased estimator for θ.

Comment on the result. [*Agra Univ. M.Sc., 1988*]

57. Let the random variables X and Y have the joint p.d.f.

$$f(x, y) = \frac{2}{\theta^2} \exp\left[-\frac{(x+y)}{\theta}\right], 0 < x < y < \infty$$

and zero elsewhere.

(*a*) Show that : $E(Y \mid x) = x + \theta$

Obtain the expected value of $X + \theta$ and compare the variance of $X + \theta$ with that of Y. [*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986*]

(*b*) Show that : $E(Y) = \frac{3}{2}\theta$, $\text{Var}(Y) = \frac{5}{4}\theta^2$.

[*Madras Univ. B.Sc., 1988*]

58. (*a*) A random sample of size n is drawn from a Poisson population with parameters λ. Obtain the minimum variance unbiased estimator of λ.

[*Delhi Univ. M.A. (Eco.), 1992*]

(*b*) Establish a necessary and sufficient condition for an unbiased estimator to be an MVU estimator.

Let $X_1, X_2, ..., X_n$ be a random sample from a Poisson distribution with parameter θ. Find an MVU estimator for $\gamma(\theta) = e^{-\theta}\theta^4/24$.

59. Define sufficiency of an estimator

Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ be the order statistics of a random sample of size 5 from the uniform distribution with p.d.f.

$$f(x, \theta) = \begin{cases} \dfrac{1}{\theta}; & 0 < x < \theta, \ 0 < \theta < \infty \\ 0, & \text{elsehwhere} \end{cases}$$

Show that $2Y_3$ is an unbiased estimator of θ. Find the conditional expectation $E[2Y_3 \mid Y_5] = \phi(Y_5)$, say. Compare the variances of $2Y_3$ and $\phi(Y_5)$.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

60. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Bernoulli population with parameter θ, $0 < \theta < 1$. Obtain a sufficient statistic for θ and show that it is complete. Hence obtain *MVU* estimator of θ.

[***Delhi Univ. B.Sc. (Stat. Hons.), 1989***]

61. Show that $T = \sum_{i=1}^{n} X_i$, is a complete sufficient statistic for the parameter θ in a random sample $X_1, X_2, \ldots, X_n$ drawn from the population with *p.d.f.*

(*a*) $$f(x, \theta) = \theta^x (1-\theta)^{1-x}\ ; x = 0, 1$$
$$= 0, \text{ elsewhere}$$

(*b*) $$f(x, \theta) = \begin{cases} e^{-\theta}\,\theta^x / x\,!\ , & x = 0, 1, 2, \ldots \\ 0 & , \text{ elsewhere} \end{cases}$$

62. If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, show that :

(*a*) $T = \bar{X}$, is complete sufficient statistic for μ, $(-\infty < \mu < \infty)$, when σ^2 is known.

(*b*) $T = \sum_{i=1}^{n} (X_i - \mu)^2$, is complete sufficient statistic for σ^2, $(0\ \sigma^2 < \infty)$, when μ is known.

15·10. Methods of Estimation. So far we have been discussing the requisites of a good estimator. Now we shall briefly outline some of the important methods for obtaining such estimators. Commonly used methods are

(*i*) *Method of Maximum Likelihood Estimation.*
(*ii*) *Method of Minimum Variance.*
(*iii*) *Method of Moments.*
(*iv*) *Method of Least Squares.*
(*v*) *Method of Minimum Chi-square*
(*vi*) *Method of Inverse Probability.*

In the following sections, we shall discuss briefly the first four methods only.

15·11. Method of Maximum Likelihood Estimation. From theoretical point of view, the most general method of estimation known is the method of *Maximum Likelihood Estimators* (M.L.E.) which was initially formulated by C.F. Gauss but as a general method of estimation was first introduced by Prof. R.A. Fisher and later on developed by him in a series of papers. Before introducing the method we will first define *Likelihood Function.*

Likelihood Function. *Definition.* Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a population with density function $f(x, \theta)$. Then the likelihood function of the sample values $x_1, x_2, \ldots, x_n$, usually denoted by $L = L(\theta)$ is their joint density function, given by

$$L = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta). \quad \ldots(15{\cdot}53)$$

L gives the relative likelihood that the random variables assume a particular set of values $x_1, x_2, \ldots, x_n$. For a given sample $x_1, x_2, \ldots, x_n, L$ becomes a function of the variable θ, the parameter.

The principle of maximum likelihood consists in finding an estimator for the unknown parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, say, which maximises the likelihood function $L(\theta)$ for variations in parameter *i.e.*, we wish to find $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ so that

$$L(\hat{\theta}) > L(\theta) \qquad \forall\ \theta \in \Theta$$

i.e., $$L(\hat{\theta}) = \text{Sup}\, L(\theta)\ \forall\ \theta \in \Theta.$$

Thus if there exists a function $\hat{\theta} = \hat{\theta}\,(x_1, x_2, \ldots, x_n)$ of the sample values which maximises L for variations in θ, then $\hat{\theta}$ is to be taken as an estimator of θ. $\hat{\theta}$ is usually called *Maximum Likelihood Estimator (M.L.E.)*. Thus $\hat{\theta}$ is the solution, if any, of

$$\frac{\partial L}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial^2 L}{\partial \theta^2} < 0 \qquad \ldots(15{\cdot}54)$$

Since $L > 0$, and $\log L$ is a non-decreasing function of L ; L and $\log L$ attain their extreme values (maxima or minima) at the same value of $\hat{\theta}$. The first of the two equations in (15·54) can be rewritten as

$$\frac{1}{L} \cdot \frac{\partial L}{\partial \theta} = 0 \quad \Rightarrow \quad \frac{\partial \log L}{\partial \theta} = 0, \qquad \ldots(15{\cdot}54a)$$

a form which is much more convenient from practical point of view.

If θ is vector valued parameter, then $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$, is given by the solution of simultaneous equations :

$$\frac{\partial}{\partial \theta_i} \log L = \frac{\partial}{\partial \theta_i} \log L\,(\theta_1, \theta_2, \ldots, \theta_k) = 0\ ;\ i = 1, 2, \ldots, k \qquad \ldots(15{\cdot}54b)$$

Equations (15·54*a*) and (15·54*b*) are usually referred to as the *Likelihood Equations* for estimating the parameters.

Remark. For the solution $\hat{\theta}$ of the likelihood equations, we have to see that the second derivative of L w.r. to θ is negative. If θ is vector valued, then for L to be maximum, the matrix of derivatives

$$\left(\frac{\partial^2 \log L}{\partial \theta_i\, \partial \theta_j}\right)_{\theta = \hat{\theta}} \text{ should be negative definite.}$$

15·11·1. Properties of Maximum Likelihood Estimators.

We make the following assumptions, known as the *Regularity Conditions* :

(*i*) The first and second order derivatives, *viz.*, $\dfrac{\partial \log L}{\partial \theta}$ and $\dfrac{\partial^2 \log L}{\partial \theta^2}$ exist and are continuous functions of θ in a range R (including the true value θ_0 of the parameter) for almost all x. For every θ in R

$$\left|\frac{\partial}{\partial \theta} \log L\right| < F_1(x) \text{ and } \left|\frac{\partial^2}{\partial \theta^2} \log L\right| < F_2(x)$$

where $F_1(x)$ and $F_2(x)$ are integrable functions over $(-\infty, \infty)$.

(*ii*) The third order derivative $\frac{\partial^2}{\partial\theta^3}\log L$ exists such that

$$\left|\frac{\partial^3}{\partial\theta^3}\cdot\log L\right| < M(x)$$

where $E[M(x)] < K$, a positive quantity.

(*iii*) For every θ in R,

$$E\left(-\frac{\partial^2}{\partial\theta^2}\log L\right) = \int_{-\infty}^{\infty}\left(-\frac{\partial^2}{\partial\theta^2}\log L\right)L\,dx$$
$$= I(\theta),$$

is finite and non-zero.

(*iv*) The range of integration is independent of θ. But if the range of integration depends on θ, then $f(x, \theta)$ vanishes at the extremes depending on θ.

This assumption is to make the differentiation under the integral sign valid.

Under the above assumptions M.L.E. possesses a number of important properties, which will be stated in the form of theorems.

Theorem 15·11. (*Cramer-Rao Theorem*). *"With probability approaching unity as $n \to \infty$, the likelihood equation $\frac{\partial}{\partial\theta}\log L = 0$, has a solution which converges in probability to the true value θ_0". In other words M.L.E.'s are consistent.*

Remark. *MLE's* are always consistent estimators but need not be unbiased. For example in sampling from $N(\mu, \sigma^2)$ population, [c.f. Example 15·31],

MLE$(\mu) = \bar{x}$ (sample mean), which is both unbiased and consistent estimator of μ.

MLE$(\sigma^2) = s^2$ (sample variance), which is consistent but not unbiased estimator of σ^2.

Theorem 15·12. (*Hazoor Bazar's Theorem*). *Any consistent solution of the likelihood equation provides a maximum of the likelihood with probability tending to unity as the sample size (n) tends to infinity.*

Theorem 15·13. (Asymptotic Normality of MLE's). *A consistent solution of the likelihood equation is asymptotically normally distributed about the true value θ_0.* Thus, $\hat{\theta}$ is asymptotically $N\left(\theta_0, \frac{I}{I(\theta_0)}\right)$ as $n \to \infty$.

Remark. Variance of M.L.E. is given by

$$V(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{1}{\left[E\left(-\frac{\partial^2}{\partial\theta^2}\log L\right)\right]} \qquad \ldots(15·55)$$

Theorem 15·14. *If M.L.E. exists, it is the most efficient in the class of such estimators.*

Theorem 15·15. *If a sufficient estimator exists, it is a function of the Maximum Likelihood Estimator.*

Proof. If $t = t(x_1, x_2, \ldots, x_n)$ is a sufficient estimator of θ, then Likelihood Function can be written as (*c.f.* Theorem 15·7)

$$L = g(t, \theta)\, h(x_1, x_2, x_3, \ldots, x_n \mid t)$$

where $g(t, \theta)$ is the density function of t and $h(x_1, x_2, \ldots, x_n \mid t)$ is the density function of the sample, given t, and is independent of θ.

$$\therefore \qquad \log L = \log g(t, \theta) + \log h(x_1, x_2, \ldots, x_n \mid t)$$

Differentiating w.r.t. θ, we get

$$\frac{\partial \log L}{\partial \theta} = \frac{\partial}{\partial \theta} \log g(t, \theta) = \psi(t, \theta), \text{ (say)}, \qquad \ldots(15·56)$$

which is a function of t and θ only.

M.L.E. is given by

$$\frac{\partial \log L}{\partial \theta} = 0 \quad \Rightarrow \quad \psi(t, \theta) = 0$$

$$\therefore \qquad \hat{\theta} = \eta(t) = \text{Some function of sufficient statistic.}$$

$$\Rightarrow \qquad \hat{t} = \psi(\theta) = \text{Some function of M.L.E.}$$

Hence the theorem.

Remark. This theorem is quite helpful in finding if a sufficient estimator exists or not.

If $\frac{\partial}{\partial \theta} \log L$ can be expressed in the form (15·56), *i.e.*, as a function of a statistic and parameter alone, then the statistic is regarded as a sufficient estimator of the parameter. If $\frac{\partial}{\partial \theta} \log L$ cannot be expressed in the form (15·56), no sufficient estimator exists in that case.

Theorem 15·16. *If for a given population with p.d.f. $f(x, \theta)$, an MVB estimator T exists for θ, then the likelihood equation will have a solution equal to the estimator T.*

Proof. Since T is an MVB estimator of θ, we have [c.f. (15·40)],

$$\frac{\partial}{\partial \theta} \log L = \frac{T - \theta}{\lambda(\theta)} = (T - \theta)\, A(\theta)$$

MLE for θ is the solution of the likelihood equation

$$\frac{\partial}{\partial \theta} \log L = 0 \quad \Rightarrow \quad \hat{\theta} = T$$

as required.

Theorem 15·17. (Invariance Property of MLE). *If T is the MLE of θ and $\psi(\theta)$ is one to one function of θ, then $\psi(T)$ is the MLE of $\psi(\theta)$.*

Example 15·31. *In random sampling from normal population $N(\mu; \sigma^2)$, find the maximum likelihood estimators for*

(i) *μ when σ^2 is known,*

(ii) *σ^2 when μ is known, and* *(iii) the simultaneous estimation of μ and σ^2.*

[Madras Univ. B.Sc. Sept., 1987]

Solution. $X \sim N(\mu, \sigma^2)$ then

$$L = \prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right\}$$

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Case (i). When σ^2 is known, the likelihood equation for estimating μ is

$$\frac{\partial}{\partial\mu}\log L = 0 \;\Rightarrow\; -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)(-1) = 0$$

or $$\sum_{i=1}^{n}(x_i-\mu) = 0 \;\Rightarrow\; \sum_{i=1}^{n}x_i - n\mu = 0$$

$\Rightarrow$ $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x} \qquad \ldots(*)$$

Hence M.L.E. for μ is the sample mean $\bar{x}$.

Case (ii). When μ is known, the likelihood equation for estimating σ^2 is

$$\frac{\partial}{\partial\sigma^2}\log L = 0 \;\Rightarrow\; -\frac{n}{2}\times\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$\Rightarrow$ $$n - \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0, \quad i.e., \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \qquad \ldots(**)$$

Case (iii). The likelihood equations for simultaneous estimation of μ and σ^2 are

$$\frac{\partial}{\partial\mu}\log L = 0 \text{ and } \frac{\partial}{\partial\sigma^2}\log L = 0, \text{ thus giving}$$

$$\hat{\mu} = \bar{x} \qquad \text{[From (*)]}$$

and $$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \qquad \text{[From (**)]}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = s^2, \text{ the sample variance.}$$

Important Note. It may be pointed out here that though

$$\left.\begin{aligned} E(\hat{\mu}) &= E(\bar{x}) = \mu \\ E(\hat{\sigma}^2) &= E(s^2) \neq \sigma^2 \end{aligned}\right\} \qquad (c.f.\ \S\ 12\cdot12)$$

Hence the maximum likelihood estimators (M.L.Es.) need not necessarily be unbiased.

Remark. Since M.L.E. is the most efficient, we conclude that in sampling from a normal population, the sample mean $\bar{x}$ is the most efficient estimator of the population mean μ.

Example 15·32. *Prove that the maximum likelihood estimate of the parameter α of a population having density function :*

$$\frac{2}{\alpha^2}(\alpha - x),\ 0 < x < \alpha$$

for a sample of unit size is $2x$, x being the sample value. Show also that the estimate is biased. **[*Burdwan Univ. B.Sc. (Maths. Hons.), 1991*]**

Solution. For a random sample of unit size ($n = 1$), the likelihood function is :

$$L(\alpha) = f(x, \alpha) = \frac{2}{\alpha^2}(\alpha - x)\ ;\ 0 < x < \alpha$$

Likelihood equation gives :

$$\frac{d}{d\alpha}\log L = \frac{d}{d\alpha}\left[\log 2 - 2\log \alpha + \log(\alpha - x)\right] = 0$$

$$\Rightarrow \quad -\frac{2}{\alpha} + \frac{1}{\alpha - x} = 0 \quad \Rightarrow \quad 2(\alpha - x) - \alpha = 0 \quad \Rightarrow \quad \alpha = 2x$$

Hence MLE of α is given by $\hat{\alpha} = 2x$.

$$E(\hat{\alpha}) = E(2X) = 2\int_0^\alpha x . f(x, \alpha)\, dx$$

$$= \frac{4}{\alpha^2}\int_0^\alpha x(\alpha - x)\, dx = \frac{4}{\alpha^2}\left|\frac{\alpha x^2}{2} - \frac{x^3}{3}\right|_0^\alpha = \frac{2}{3}\alpha$$

Since $E(\hat{\alpha}) \neq \alpha$, $\hat{\alpha} = 2x$ is not an unbiased estimate of α.

Example 15·33. (*a*) *Find the maximum likelihood estimate for the parameter λ of a Poisson distribution on the basis of a sample of size n. Also find its variance.*

(*b*) *Show that the sample mean $\bar{x}$, is sufficient for estimating the parameter λ of the Poisson distribution.*

Solution. The probability function of the Poisson distribution with parameter λ is given by

$$P(X = x) = f(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{x\,!}\ ;\ x = 0, 1, 2, \ldots$$

Likelihood function of random sample $x_1, x_2, \ldots, x_n$ of n observations from this population is

$$L = \prod_{i=1}^{n} f(x_i, \lambda) = \frac{e^{-n\lambda}\,\lambda^{\sum_{i=1}^{n} x_i}}{x_1\,!\ x_2\,! \ldots x_n\,!}$$

$$\therefore \quad \log L = -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\log\lambda - \sum_{i=1}^{n}\log(x_i\,!)$$

$$= -n\lambda + n\bar{x}\log\lambda - \sum_{i=1}^{n}\log(x_i\,!)$$

The likelihood equation for estimating λ is

$$\frac{\partial}{\partial\lambda}\log L = 0 \quad \Rightarrow \quad -n + \frac{n\bar{x}}{\lambda} = 0 \quad \Rightarrow \quad \lambda = \bar{x}$$

Thus the M.L.E. for λ is the sample mean $\bar{x}$.
The variance of the estimate is given by

$$\frac{1}{V(\hat{\lambda})} = E\left[-\frac{\partial^2}{\partial\lambda^2}(\log L)\right] \qquad [c.f.\ (15\cdot55)]$$

$$= E\left[-\frac{\partial}{\partial\lambda}\left(-n + \frac{n\bar{x}}{\lambda}\right)\right] = E\left[-\left(-\frac{n\bar{x}}{\lambda^2}\right)\right] = \frac{n}{\lambda^2}E(\bar{x}) = \frac{n}{\lambda}$$

$$\therefore \quad V(\hat{\lambda}) = \lambda/n$$

(*b*) For the Poisson distribution with parameter λ, we have

$$\frac{\partial}{\partial\lambda}\log L = -n + \frac{n\bar{x}}{\lambda}$$

$$= n\left(\frac{\bar{x}}{\lambda} - 1\right) = \psi(\bar{x}, \lambda), \text{ a function of } \bar{x} \text{ and } \lambda \text{ only.}$$

Hence (*c.f.* Remark Theorem 15·15), $\bar{x}$ is sufficient for estimating λ.

Example 15·34. *Let $x_1, x_2, \ldots, x_n$ denote random sample of size n from a uniform population with p.d.f.*

$$f(x, \theta) = 1\ ;\ \theta - \frac{1}{2} \le x \le \theta + \frac{1}{2}, -\infty < \theta < \infty$$

Obtain M.L.E. for θ. **[*Delhi Univ. M.C.A., 1987*]**

Solution. Here

$$L = L(\theta; x_1, x_2, \ldots, x_n) = 1,\ \theta - \frac{1}{2} \le x_i \le \theta + \frac{1}{2}$$
$$= 0, \text{ elsewhere}$$

If $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is the ordered sample then

$$\theta - \frac{1}{2} \le x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} \le \theta + \frac{1}{2}$$

Thus L attains the maximum if

$$\theta - \frac{1}{2} \le x_{(1)} \quad \wedge \quad x_{(n)} \le \theta + \frac{1}{2}$$

$$\Rightarrow \quad \theta \le x_{(1)} + \frac{1}{2} \ \wedge \ x_{(n)} - \frac{1}{2} \le \theta$$

Hence *every statistic $t = t(x_1, x_2, \ldots, x_n)$ such that*

$$x_{(n)} - \frac{1}{2} \le t\,(x_1, x_2, \ldots, x_n) \le x_{(1)} + \frac{1}{2}$$

provides an M.L.E. for θ.

Remark. This example illustrates that M.L.E. for a parameter need not be unique.

Example 15·35. *Find the M.L.E. of the parameters* α *and* λ, (λ *being large*), *of the distribution :*

$$f(x;\, \alpha, \lambda) = \frac{1}{\Gamma(\lambda)} \left(\frac{\lambda}{\alpha} \right)^{\lambda} e^{-\lambda x/\alpha}\, x^{\lambda - 1};\; 0 \le x < \infty,\ \lambda > 0$$

You may use that for large values of λ,

$$\psi(\lambda) = \frac{\partial}{\partial \lambda} \log \Gamma(\lambda) = \log \lambda - \frac{1}{2\lambda}$$

and $$\psi'(\lambda) = \frac{1}{\lambda} + \frac{1}{2\lambda^2}$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1985*]

Solution. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from the given population. Then

$$L = \prod_{i=1}^{n} f(x_i;\, \alpha, \lambda) = \left(\frac{1}{\Gamma(\lambda)} \right)^{n} \cdot \left(\frac{\lambda}{\alpha} \right)^{n\lambda} \cdot \exp\left[-\frac{\lambda}{\alpha} \sum_{i=1}^{n} x_i \right] \cdot \prod_{i=1}^{n} (x_i^{\lambda-1})$$

$$\therefore \quad \log L = -n \log \Gamma(\lambda) + n\lambda(\log \lambda - \log \alpha) - \frac{\lambda}{\alpha} \sum_{i=1}^{n} x_i + (\lambda - 1) \sum_{i=1}^{n} \log x_i$$

If G is the geometric mean of $x_1, x_2, \ldots, x_n$, then

$$\log G = \frac{1}{n} \sum_{i=1}^{n} \log x_i \quad \Rightarrow \quad n \log G = \sum_{i=1}^{n} \log x_i$$

$$\therefore \quad \log L = -n \log \Gamma(\lambda) + n\lambda\,(\log \lambda - \log \alpha) - \frac{\lambda}{\alpha} n\bar{x} + (\lambda - 1).\, n \log G$$

where G is independent of λ and α.

The likelihood equations for the simultaneous estimation of α and λ are :

$$\frac{\partial}{\partial \alpha} \log L = 0 \;\ldots(1) \quad \text{and} \quad \frac{\partial}{\partial \lambda} \log L = 0 \qquad \ldots(2)$$

(1) gives

$$-\frac{n\lambda}{\alpha} + \frac{\lambda}{\alpha^2} \cdot n\bar{x} = 0 \;\Rightarrow\; -1 + \frac{\bar{x}}{\alpha} = 0 \;\Rightarrow\; \hat{\alpha} = \bar{x} \qquad \ldots(*)$$

(2) gives (for large values of λ),

$$-n\left(\log \lambda - \frac{1}{2\lambda} \right) + n\left[1.(\log \lambda - \log \alpha) + \lambda.\, \frac{1}{\lambda} \right] - \frac{n\bar{x}}{\alpha} + n \log G = 0$$

$$\Rightarrow \quad \frac{1}{2\lambda} + \left(1 - \log \alpha + \log G - \frac{\bar{x}}{\alpha} \right) = 0$$

$$\Rightarrow \quad 1 + 2\lambda\,(\log G - \log \bar{x}) = 0 \qquad \text{[From (*)]}$$

$$\Rightarrow \qquad 1 - 2\lambda \log\left(\frac{\bar{x}}{G}\right) = 0, \quad i.e., \quad \hat{\lambda} = \frac{1}{2\log(\bar{x}/G)}$$

Hence the M.L.E.s for α and λ are given by

$$\hat{\alpha} = \bar{x} \quad \text{and} \quad \hat{\lambda} = \frac{1}{2\log(\bar{x}/G)}\,,$$

Example 15·36. *In sampling from a power series distribution with p.d.f.*

$$f(x, \theta) = a_x \theta^x / \psi(\theta)\ ; \ x = 0, 1, 2, \ldots$$

where a_x may be zero for some x, show that MLE of θ is a root of the equation

$$\bar{X} = \frac{\theta\,\psi'(\theta)}{\psi(\theta)} = \mu(\theta), \qquad \ldots(*)$$

where $\mu(\theta) = E(X)$.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

Solution. Likelihood function is given by :

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \prod_{i=1}^{n}\left[\frac{a_{x_i}\,\theta^{x_i}}{\psi(\theta)}\right] = \left[\prod_{i=1}^{n} a_{x_i}\right]\frac{\theta^{\Sigma x_i}}{[\psi(\theta)]^n}$$

$$\Rightarrow \qquad \log L = \sum_{i=1}^{n} \log a_{x_i} + \log\theta \cdot \sum_{i=1}^{n} x_i - n\log\psi(\theta)$$

Likelihood equation for estimating θ gives :

$$\frac{\partial}{\partial\theta}\log L = 0 = \frac{\Sigma x_i}{\theta} - \frac{n\psi'(\theta)}{\psi(\theta)}$$

$$\Rightarrow \qquad \bar{X} = \frac{\Sigma x_i}{n} = \frac{\theta\,\psi'(\theta)}{\psi(\theta)} = \mu(\theta), \text{ (say)}. \qquad \ldots(*)$$

Hence MLE of θ is a root of equation (*).

We have :

$$E(X) = \sum_{x=0}^{\infty} x f(x, \theta) = \sum_{x=0}^{\infty}\left[x\left\{\frac{a_x\,\theta^x}{\psi(\theta)}\right\}\right] \qquad \ldots(**)$$

$$\sum_{x=0}^{\infty} f(x, \theta) = 1 \quad \Rightarrow \quad \sum_{x=0}^{\infty}\frac{a_x\,\theta^x}{\psi(\theta)} = 1 \quad \Rightarrow \quad \sum_{x=0}^{\infty} a_x\,\theta^x = \psi(\theta)$$

Differentiating w.r. to θ, we get

$$\sum_x [a_x \cdot x\theta^{x-1}] = \psi'(\theta)$$

$$\Rightarrow \qquad \sum_x\left[a_x \cdot \frac{x\,\theta^x}{\psi(\theta)}\right] = \frac{\theta.\psi'(\theta)}{\psi(\theta)}$$

$$\Rightarrow \qquad E(X) = \mu(\theta) = \bar{X}, \qquad \text{[From (**) and (*)]}$$

as required.

Example 15·37. *(a) Let $x_1, x_2, \ldots, x_n$ be a random sample from the uniform distribution with p.d.f.*

$$f(x, \theta) = \frac{1}{\theta},\ 0 < x < \infty,\ \theta > 0$$
$$= 0,\ \text{elsewhere}$$

Obtain the maximum likelihood estimator for θ.

[*Lucknow Univ. B.Sc., 1992*]

(*b*) *Obtain the M.L.Es. for α and β for the rectangular population*

$$f(x; \alpha, \beta) = \begin{cases} \dfrac{1}{\beta - \alpha}, & \alpha < x < \beta \\ 0, & \text{elsewhere} \end{cases}$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989; Gujarat Univ. B.Sc. 1992*]

Solution. (*a*) Here

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \frac{1}{\theta} \cdot \frac{1}{\theta} \cdots \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n \qquad \ldots(*)$$

Likelihood equation, *viz.*, $\frac{\partial}{\partial \theta} \log L = 0$, gives

$$\frac{\partial}{\partial \theta}(-n \log \theta) = 0 \quad \Rightarrow \quad \frac{-n}{\theta} = 0 \quad \Rightarrow \quad \hat{\theta} = \infty,$$

obviously an absurd result.

In this case we locate M.L.E. as follows :

We have to choose θ so that L in (*) is maximum. Now L is maximum if θ is minimum.

Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the *ordered* sample of n independent observations from the given population so that

$$0 \le x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} \le \theta \quad \Rightarrow \quad \theta \ge x_{(n)}$$

Since the minimum value of θ consistent with the sample is $x_{(n)}$, the largest sample observation, $\hat{\theta} = x_{(n)}$.

∴ M.L.E. for $\theta = x_{(n)}$ = The largest sample observation.

(*b*) Here

$$L = \left(\frac{1}{\beta - \alpha}\right)^n \qquad \ldots(**)$$

∴ $\log L = -n \log (\beta - \alpha)$

The likelihood equations for α and β give

$$\left.\begin{aligned} \frac{\partial}{\partial \alpha} \log L &= 0 = \frac{n}{\beta - \alpha} \\ \text{and} \quad \frac{\partial}{\partial \beta} \log L &= 0 = \frac{-n}{\beta - \alpha} \end{aligned}\right\}$$

Each of these equations gives $\beta - \alpha = \infty$, an obviously negative result. So, we find M.L.Es for α and β by some other means.

Now L in (**) is maximum if $(\beta - \alpha)$ is minimum, *i.e.*, if β takes the minimum possible value and α takes the maximum possible value.

As in part (*a*), if $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is an *ordered* random sample from this population, then $\alpha \le x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} \le \beta$. Thus $\beta \ge x_{(n)}$ and $\alpha \le x_{(1)}$. Hence the minimum possible value of β consistent with the sample is $x_{(n)}$ and the maximum possible value of α consistent with the sample is $x_{(1)}$. Hence L is maximum if $\beta = x_{(n)}$ and $\alpha = x_{(1)}$.

$\therefore$ M.L.E. for α and β are given by

$$\hat{\alpha} = x_{(1)} = \text{The smallest sample observation}$$

and
$$\hat{\beta} = x_{(n)} = \text{The largest sample observation.}$$

Example 15·38. *State as precisely as possible the properties of the M.L.E. Obtain the M.L.Es. of α and β for a random sample from the exponential population*

$$f(x\,;\,\alpha, \beta) = y_0\, e^{-\beta\,(x-\alpha)}, \alpha \le x \le \infty, \beta > 0$$

y_0 being a constant.

Solution. Here first of all we shall determine the constant y_0 from the consideration that the total area under a probability curve is unity.

$$\therefore \qquad y_0 \int_{\alpha}^{\infty} \exp\left[-\beta(x-\alpha)\right] dx = 1$$

or
$$y_0 \left| \frac{e^{-\beta(x-\alpha)}}{-\beta} \right|_{\alpha}^{\infty} = 1 \quad \Rightarrow \quad -\frac{y_0}{\beta}(0-1) = 1 \quad \Rightarrow \quad y_0 = \beta$$

$$\therefore \qquad f(x\,;\,\alpha, \beta) = \beta\, e^{-\beta(x-\alpha)}, \alpha \le x < \infty$$

If $x_1, x_2, \ldots, x_n$ is a random sample of n observations from this population, then

$$L = \prod_{i=1}^{n} f(x_i\,;\,\alpha, \beta) = \beta^n \exp\left\{-\beta \sum_{i=1}^{n} (x_i - \alpha)\right\} = \beta^n e^{-n\beta(\bar{x}-\alpha)}$$

$$\therefore \qquad \log L = n \log \beta - n\beta(\bar{x} - \alpha) \qquad \ldots(*)$$

The likelihood equations for estimating α and β give

$$\frac{\partial}{\partial \alpha} \log L = 0 = n\beta \qquad \ldots(**)$$

and
$$\frac{\partial}{\partial \beta} \log L = 0 = \frac{n}{\beta} - n(\bar{x} - \alpha) \qquad \ldots(***)$$

Equation (**) gives $\beta = 0$, which is obviously inadmissible and this on substitution in (***) gives $\alpha = \infty$, a nugatory result. Thus the likelihood equations fail to give us valid estimates of α and β and we try to locate M.L.Es. for α and β by maximising L directly.

$$L \text{ is maximum} \quad \Rightarrow \quad \log L \text{ is maximum.}$$

From (*), $\log L$ is maximum (for any value of β), if $(\bar{x} - \alpha)$ is minimum, which is so if α is maximum.

If $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is ordered sample from this population then

$$\alpha \le x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} < \infty,$$

so that the maximum value of α consistent with the sample is $x_{(1)}$, the smallest sample observation, *i.e.*,

$$\hat{\alpha} = x_{(1)}$$

Consequently, (***) gives

$$\frac{1}{\beta} = \bar{x} - \hat{\alpha} = \bar{x} - x_{(1)} \quad \Rightarrow \quad \hat{\beta} = \frac{1}{\bar{x} - x_{(1)}}$$

Hence M.L.Es. for α and β are given by

$$\hat{\alpha} = x_{(1)} \text{ and } \hat{\beta} = \frac{1}{\bar{x} - x_{(1)}}$$

Remarks 1. Whenever the given probability function involves a constant and the range of the variable is dependent on the parameter(s) to be estimated, first of all we should determine the constant by taking the total probability as unity and then proceed with the estimation part.

2. From the last two examples, it is obvious that whenever the range of the variable involves the parameter(s) to be estimated, the likelihood equations fail to give us valid estimates and in this case M.L.Es are obtained by adopting some other approach of maximising L or $\log L$ directly.

Example 5·39. *Obtain the maximum likelihood estimate of θ in*

$$f(x, \theta) = (1 + \theta)\, x^{\theta}, 0 < x < 1,$$

based on an independent sample of size n. Examine whether this estimate is sufficient for θ.

Solution.

$$L(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = (1+\theta)^n \cdot \left(\prod_{i=1}^{n} x_i\right)^{\theta}$$

$$\Rightarrow \quad \log L = n \log (1 + \theta) + \theta . \sum_{i=1}^{n} \log x_i$$

$$\frac{\partial}{\partial \theta} \log L = \frac{n}{1+\theta} + \sum_{i=1}^{n} \log x_i = 0$$

$$\Rightarrow \quad n + \theta \sum_i \log x_i + \sum_i \log x_i = 0$$

$$\therefore \quad \hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log x_i} - 1 = \frac{-n}{\log\left(\prod_{i=1}^{n} x_i\right)} - 1 \quad \ldots(*)$$

Also $$L(x, \theta) = \left\{(1+\theta)^n \cdot \left(\prod_{i=1}^{n} x_i\right)^{\theta - 1}\right\} . \left(\prod_{i=1}^{n} x_i\right)$$

Hence by Factorisation theorem, $T = \left(\prod_{i=1}^{n} x_i\right)$ is a sufficient statistic for θ, and $\hat{\theta}$ being a one to one function of sufficient statistic $\left(\prod_{i=1}^{n} x_i\right)$, is also sufficient for θ.

Example 15·40. *(a) Obtain the most general form of distribution differentiable in θ, for which the sample mean is the M.L.E.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1983]

(b) Show that the most general continuous distribution for which the M.L.E. of a parameter θ is the geometric mean of the sample is

$$f(x, \theta) = \left(\frac{x}{\theta}\right)^{\theta \cdot \frac{\partial \psi}{\partial \theta}} \exp\left[\psi(\theta) + \xi(x)\right],$$

where $\psi(\theta)$ and $\xi(x)$ are arbitrary functions of θ and x respectively.

Solution. (*a*) We have $L = \prod_{i=1}^{n} f(x_i, \theta)$

$$\Rightarrow \qquad \log L = \sum_{i=1}^{n} \log f(x_i, \theta) = \sum_{x} \log f, \qquad [f = f(x, \theta)]$$

the summation extending to all the values of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ in the sample. The likelihood equation is

$$\frac{\partial}{\partial \theta} \log L = 0, \quad i.e., \quad \frac{\partial}{\partial \theta}\left(\sum_{x} \log f\right) = 0$$

$$\Rightarrow \qquad \sum_{x} \frac{\partial}{\partial \theta} \log f = 0 \quad \Rightarrow \quad \sum_{x} \frac{1}{f} \cdot \frac{\partial f}{\partial \theta} = 0 \qquad \ldots(*)$$

W are given that the solution of (*) is

$$\theta = \frac{1}{n} \sum x \quad \text{or} \quad n\theta = \sum x$$

$$\Rightarrow \qquad \sum_{x} (x - \theta) = 0 \qquad \ldots(**)$$

Since this is true for all values of x and θ, we get from (*) and (**),

$$\frac{1}{f} \cdot \frac{\partial f}{\partial \theta} = A(x - \theta),$$

where A is independent of x but may be function of θ. Let us take

$$A = \frac{\partial^2 \psi}{\partial \theta^2} \text{ where } \psi = \psi(\theta) \text{ is any arbitrary function of } \theta.$$

Thus
$$\frac{\partial}{\partial \theta} \log f = \frac{\partial^2 \psi}{\partial \theta^2} (x - \theta)$$

Integrating w.r. to θ (partially), we get

$$\log f = (x - \theta) \cdot \frac{\partial \psi}{\partial \theta} - \int \frac{\partial \psi}{\partial \theta} (-1)\, d\theta + \xi(x) + k$$

where $\xi(x)$ is an arbitrary function of x and k is arbitrary constant.

$$\therefore \qquad \log f = (x-\theta)\,.\,\frac{\partial\psi}{\partial\theta} + \psi(\theta) + \xi(x) + k$$

Hence $$f = \text{Const. } \exp\left[(x-\theta)\,\frac{\partial\psi}{\partial\theta} + \psi(\theta) + \xi(x)\right]$$

which is the probability function of the required distribution.

Remark. In particular, if we take.

$$\psi(\theta) = \frac{\theta^2}{2} \quad \text{and} \quad \xi(x) = -\frac{x^2}{2}, \text{ then}$$

$$f = \text{Const. } \exp\left[(x-\theta)\,.\,\theta + \frac{\theta^2}{2} - \frac{x^2}{2}\right]$$

$$= \text{Const. } \exp\left[-\tfrac{1}{2}(x^2 + \theta^2 - 2\theta x)\right]$$

$$= \text{Const. } \exp\left\{-\tfrac{1}{2}(x-\theta)^2\right\}$$

which is the probability function of the normal distribution with mean θ and unit variance.

(*b*) Here the solution of the likelihood equation

$$\frac{\partial}{\partial\theta}\log L = \sum_x \frac{\partial}{\partial\theta}\log f = 0 \qquad \ldots(*)$$

is $$\theta = (x_1, x_2, \ldots, x_n)^{1/n}$$

$$\Rightarrow \qquad \log\theta = \frac{1}{n}\sum_x \log x \quad \Rightarrow \quad \sum_x(\log x - \log\theta) = 0 \qquad \ldots(**)$$

Since this is true for all x and all θ, we get from (*) and (**)

$$\frac{\partial}{\partial\theta}\log f = (\log x - \log\theta)\,A(\theta)$$

where $A(\theta)$ is an arbitrary function of θ and is independent of x.

Integrating w.r. to θ (partially), we get

$$\log f = \log x \int A(\theta)\,d\theta - \int A(\theta)\log\theta\,d\theta + \xi(x)$$

where $\xi(x)$ is an arbitrary function of x alone.

If we take $\int A(\theta)\,d\theta = A_1(\theta)$, then

$$\log f = \log x\,.\,A_1(\theta) - \left[A_1(\theta)\log\theta - \int A_1(\theta)\,.\,\frac{1}{\theta}\,d\theta\right] + \xi(x)$$

$$= A_1(\theta)\log(x/\theta) + \int\frac{A_1(\theta)}{\theta}\,d\theta + \xi(x)$$

Let us take

$$A_1(\theta) = \theta\,\frac{\partial\psi}{\partial\theta}, \text{ (suggested by the answer)}$$

where $\psi = \psi(\theta)$ is an arbitrary function of θ alone.

$$\therefore \qquad \log f = \theta\,\frac{\partial\psi}{\partial\theta}\log(x/\theta) + \int\frac{\partial\psi}{\partial\theta}\,d\theta + \xi(x)$$

$$= \theta \frac{\partial \psi}{\partial \theta} . \log (x/\theta) + \psi(\theta) + \xi(x)$$

$$= \log \left[\left(\frac{x}{\theta} \right)^{\theta \frac{\partial \psi}{\partial \theta}} \right] + \psi(\theta) + \xi(x)$$

Hence $\quad f = f(x, \theta) = \left(\frac{x}{\theta} \right)^{\theta \frac{\partial \psi}{\partial \theta}} . \exp [\psi(\theta) + \xi(x)].$

Example 15·41. *A sample of size n is drawn from each of the four normal populations which have the same variance σ^2. The means of the four populations are $a + b + c$, $a + b - c$, $a - b + c$ and $a - b - c$. What are the M.L.Es. for a, b, c, and σ^2?*

Solution. Let the sample observations be denoted by x_{ij}, $i = 1, 2, 3, 4$; $j = 1, 2, \ldots, n$. Since the four samples, from the four normal populations are independent, the likelihood function L of all the sample observations x_{ij}, $(i = 1, 2, 3, 4; j = 1, 2, \ldots, n)$, is given by

$$L = \left(\frac{1}{\sqrt{2\pi}\, \sigma} \right)^{4n} . \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{4} \sum_{j=1}^{n} (x_{ij} - \mu_i)^2 \right\}$$

where μ_i, $(i = 1, 2, 3, 4)$ is mean of the ith population.

$$\therefore \quad L = \left(\frac{1}{\sqrt{2\pi}\, \sigma} \right)^{4n} . \exp \left[-\frac{1}{2\sigma^2} \left\{ \sum_j (x_{1j} - \mu_1)^2 + \sum_j (x_{2j} - \mu_2)^2 + \sum_j (x_{3j} - \mu_3)^2 + \sum_j (x_{4j} - \mu_4)^2 \right\} \right]$$

$$\therefore \quad \log L = k - 2n \log \sigma^2 - \frac{1}{2\sigma^2} \left[\sum_j (x_{1j} - a - b - c)^2 + \sum_j (x_{2j} - a - b + c)^2 + \sum_j (x_{3j} - a + b - c)^2 + \sum_j (x_{4j} - a + b + c)^2 \right]$$

where k is a constant w.r. to a, b, c and σ^2.

The M.L.Es. for a, b, c and σ^2 are the solutions of the simultaneous equations (maximum likelihood equations for estimating a, b, c and σ^2) :

$$\frac{\partial}{\partial a} \log L = 0 \quad \ldots(1) \qquad \frac{\partial}{\partial b} \log L = 0 \quad \ldots(2)$$

$$\frac{\partial}{\partial c} \log L = 0 \quad \ldots(3) \qquad \frac{\partial}{\partial \sigma^2} \log L = 0 \quad \ldots(4)$$

(1) gives

$$-\frac{1}{2\sigma^2} \left[\sum_j (x_{1j} - a - b - c)(-2) + \sum_j (x_{2j} - a - b + c)(-2) + \sum_j (x_{3j} - a + b - c)(-2) + \sum_j (x_{4j} - a + b + c)(-2) \right] = 0$$

$$\Rightarrow \quad \sum_j (x_{1j} + x_{2j} + x_{3j} + x_{4j})$$

$$+n\,[((-a-b-c)+(-a-b+c)+(-a+b-c)+(-a+b+c)] = 0$$

$$\Rightarrow \quad \sum_{j=1}^{n} \left(\sum_{i=1}^{4} x_{ij} \right) + n(-4a) = 0.$$

$$\therefore \quad \hat{a} = \frac{1}{4n} \sum_{i=1}^{4} \sum_{j=1}^{n} x_{ij} = \bar{x}$$

Now (2) gives

$$-\frac{1}{2\sigma^2} \Big[\sum_j (x_{1j} - a - b - c)(-2) + \sum_j (x_{2j} - a - b + c)(-2)$$

$$+ \sum_j (x_{3j} - a + b - c)(2) + \sum_j (x_{4j} - a + b + c)(2) \Big] = 0$$

$$\Rightarrow \sum_j x_{1j} + \sum_j x_{2j} - \sum_j x_{3j} - \sum_j x_{4j}$$

$$+n[(-a-b-c) + (-a-b+c) - (-a+b-c) - (-a+b+c)] = 0$$

$$\Rightarrow \quad \sum x_{1j} + \sum x_{2j} - \sum x_{3j} - \sum x_{4j} - 4nb = 0$$

$$\therefore \quad \hat{b} = \frac{1}{4} \left[\frac{1}{n} \sum x_{1j} + \frac{1}{n} \sum x_{2j} - \frac{1}{n} \sum x_{3j} - \frac{1}{n} \sum x_{4j} \right]$$

$$\Rightarrow \quad \hat{b} = (\bar{x}_1 + \bar{x}_2 - \bar{x}_3 - \bar{x}_4)/4,$$

where $\bar{x}_i$ is the mean of the ith sample.

Similarly (3) will give

$$\hat{c} = \frac{1}{4} (\bar{x}_1 - \bar{x}_2 + \bar{x}_3 - \bar{x}_4)/4$$

Equation (4) gives

$$-\frac{2n}{\sigma^2} + \frac{1}{2\sigma^4} \Big[\sum_j (x_{1j} - a - b - c)^2 + \sum_j (x_{2j} - a - b + c)^2$$

$$+ \sum_j (x_{3j} - a + b - c)^2 + \sum_j (x_{4j} - a + b + c)^2 \Big] = 0$$

$$\therefore \quad \hat{\sigma}^2 = \frac{1}{4n} \Big[\sum_j (x_{1j} - \hat{a} - \hat{b} - \hat{c})^2 + \sum_j (x_{2j} - \hat{a} - \hat{b} + \hat{c})^2$$

$$+ \sum_j (x_{3j} - \hat{a} + \hat{b} - \hat{c})^2 + \sum_j (x_{4j} - \hat{a} + \hat{b} + \hat{c})^2 \Big]$$

Example 15·42. *The following table gives probabilities and observed frequencies in four classes AB Ab, aB and ab in a genetical experiment. Estimate the parameter θ by the method of maximum likelihood and find its standard error.*

Class	Probability	Observed frequency
AB	$\frac{1}{4}(2+\theta)$	108
Ab	$\frac{1}{4}(1-\theta)$	27
aB	$\frac{1}{4}(1-\theta)$	30
ab	$\frac{1}{4}\theta$	8

Solution. Using multinomial probability law, we have

$$L = L(\theta) = \frac{n\,!}{n_1\,!\;n_2\,!\;n_3\,!\;n_4\,!}\,p_1^{n_1}\,p_2^{n_2}\,p_3^{n_3}\,p_4^{n_4},\ \Sigma p_i = 1,\ \Sigma n_i = n$$

$$\Rightarrow \quad \log L = C + n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 \log p_4,$$

where $\quad C = \log\left[\dfrac{n\,!}{n_1\,!\;n_2\,!\;n_3\,!\;n_4\,!}\right]$, is a constant.

$$\therefore \quad \log L = C + n_1 \log\left(\frac{2+\theta}{4}\right) + n_2 \log\left(\frac{1-\theta}{4}\right) + n_3 \log\left(\frac{1-\theta}{4}\right) + n_4 \log\left(\frac{\theta}{4}\right)$$

Likelihood equation gives :

$$\frac{\partial \log L}{\partial \theta} = \frac{n_1}{2+\theta} - \frac{n_2}{1-\theta} - \frac{n_3}{1-\theta} + \frac{n_4}{\theta} = 0 \qquad \ldots(*)$$

$$\Rightarrow \qquad \frac{n_1}{2+\theta} - \frac{(n_2+n_3)}{1-\theta} + \frac{n_4}{\theta} = 0$$

Taking $n_1 = 108$, $n_2 = 27$, $n_3 = 30$ and $n_4 = 8$, we get

$$\frac{108}{2+\theta} - \frac{(27+30)}{1-\theta} + \frac{8}{\theta} = 0$$

$$\Rightarrow \quad 108\theta\,(1-\theta) - 57\theta(2+\theta) + 8(1-\theta)(2+\theta) = 0$$

$$\Rightarrow \quad 173\,\theta^2 + 14\theta - 16 = 0$$

$$\Rightarrow \quad \theta = \frac{-14 \pm \sqrt{196 + 11072}}{346} = -0{\cdot}34 \text{ and } 0{\cdot}26$$

But θ, being the probability cannot be negative. Hence M.L.E. of θ is given by

$$\hat{\theta} = 0{\cdot}26 \qquad \ldots(**)$$

Differentiating (*) again partially w.r. to θ, we get

$$\frac{\partial^2 \log L}{\partial \theta^2} = \frac{-n_1}{(2+\theta)^2} - \frac{(n_2+n_3)}{(1-\theta)^2} - \frac{n_4}{\theta^2}$$

$$-E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) = \frac{E(n_1)}{(2+\theta)^2} + \frac{E(n_2)+E(n_3)}{(1-\theta)^2} + \frac{E(n_4)}{\theta^2}$$

$$= \frac{np_1}{(2+\theta)^2} + \frac{n(p_2+p_3)}{(1-\theta)^2} + \frac{np_4}{\theta^2}$$

$$= \frac{n(2+\theta)}{4(2+\theta)^2} + \frac{n(1-\theta)}{2(1-\theta)^2} + \frac{n\theta}{4\theta^2}$$

$$\therefore \qquad I(\theta) = \frac{n}{4(2+\hat{\theta})} + \frac{n}{2(1-\hat{\theta})} + \frac{n}{4\hat{\theta}};\ n = \Sigma n_i = 173.$$

$$= 173\left[\frac{1}{4 \times 2{\cdot}26} + \frac{1}{2 \times 0{\cdot}74} + \frac{1}{4 \times 0{\cdot}26}\right]$$

$$= 173\ [0{\cdot}11 + 0{\cdot}67 + 0{\cdot}96] = 173 \times 1{\cdot}74 = 301{\cdot}02$$

$$\text{S.E.}(\hat{\theta}) = \sqrt{1/I(\theta)} = \frac{1}{\sqrt{301{\cdot}02}} = 0{\cdot}0576$$

[*c.f.* (15·55) Theorem 15·13)]

15·12. Method of Minimum Variance. [Minimum Variance Unbiased Estimates (M.V.U.E.)]. In this section we shall look for estimates which (*i*) are unbiased and (*ii*) have minimum variance.

If $L = \prod_{i=1}^{n} f(x_i, \theta)$, is the likelihood function of a random sample of n observations $x_1, x_2, \ldots, x_n$ from a population with probability function $f(x, \theta)$, then the problem is to find a statistic $t = t\ (x_1, x_2, \ldots, x_n)$, such that

$$E(t) = \int_{-\infty}^{\infty} t.L\ d\mathbf{x} = \gamma(\theta) \quad \Rightarrow \quad \int_{-\infty}^{\infty} [t - \gamma(\theta)]\ L\ d\mathbf{x} = 0 \qquad \ldots(15{\cdot}57)$$

$$\text{and } V(t) = \int_{-\infty}^{\infty} [t - E(t)]^2\ L\ d\mathbf{x} = \int_{-\infty}^{\infty} [t - \gamma(\theta)]^2\ L\ d\mathbf{x} \qquad \ldots(15{\cdot}58)$$

is minimum, where

$\int_{-\infty}^{\infty} d\mathbf{x}$ represents the n-fold integration

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} dx_1\ dx_2 \ldots dx_n$$

In other words, we have to minimise (15·58) subject to the condition (15·57).

For detailed discussion of this method see *MVU* Estimators (§ 15·5·2) and Cramer-Rao Inequality (§ 15·7).

15·13. Method of Moments. This method was discovered and studied in detail by Karl Pearson.

Let $f(x\ ;\ \theta_1, \theta_2, \ldots, \theta_k)$ be the density function of the parent population with k parameters $\theta_1, \theta_2, \ldots, \theta_k$. If μ'_r denotes the rth moment about origin, then

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x\ ;\ \theta_1, \theta_2, \ldots, \theta_k)\ dx, \quad (r = 1, 2, \ldots, k) \qquad \ldots(15{\cdot}59)$$

In general $\mu_1', \mu_2', \ldots, \mu_k'$ will be functions of the parameters $\theta_1, \theta_2, \ldots, \theta_k$.

Let x_i, $i = 1, 2, \ldots, n$ be a random sample of size n from the given population. The method of moments consists in solving the k-equations (15·59) for $\theta_1, \theta_2, \ldots, \theta_k$ in terms of $\mu_1', \mu_2', \ldots, \mu_k'$ and then replacing these moments μ_r'; $r = 1, 2, \ldots, k$ by the sample moments.

e.g., $\hat{\theta}_i = \theta_i(\hat{\mu}_1', \hat{\mu}_2', \ldots, \hat{\mu}_k')$

$= \theta_i(m_1', m_2', \ldots, m_k')\ ;\ i = 1, 2, \ldots, k$

where m_i is the ith moment about origin in the sample.

Then by the method of moments $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ are the required estimators of $\theta_1, \theta_2, \ldots, \theta_k$ respectively.

Remarks. 1. Let $(x_1, x_2, \ldots, x_n)$ be a random sample of size n from a population with p.d.f. $f(x, \theta)$. Then X_i, $(i = 1, 2, \ldots, n)$ are i.i.d. $\Rightarrow X_i^r$, $(i = 1, 2, \ldots, n)$ are i.i.d r.v's. Hence if $E(X_i^r)$ exists, then by W.L.L.N., we get

$$\frac{1}{n}\sum_{i=1}^{n} x_i^r \xrightarrow[p]{p} E(X_1^r)$$

$$\Rightarrow \qquad m_r' \longrightarrow \mu_r' \qquad \ldots(15\cdot60)$$

Hence the sample moments are consistent estimators of the corresponding population moments.

2. It has been shown that under quite general conditions, the estimates obtained by the method of moments are asymptotically normal but not, in general, efficient.

3. Generally the method of moments yields less efficient estimators than those obtained from the principle of maximum likelihood. The estimators obtained by the method of moments are identical with those given by the method of maximum likelihood if the probability mass function or probability density function is of the form

$$f(x, \theta) = \exp\,[b_0 + b_1 x + b_2 x^2 + \ldots] \qquad \ldots(15\cdot61)$$

where b's are independent of x but may depend on $\theta = (\theta_1, \theta_2, \ldots)$.

(15·61) implies that

$$L(x_1, x_2, \ldots, x_n\,;\,\theta) = \exp\,[nb_0 + b_1\sum x_i + b_2 \sum x_i^2 + \ldots]$$

$$\Rightarrow \qquad \frac{\partial}{\partial\theta_j}\log L = a_0 + a_1\sum x_i + a_2\sum x_i^2 + a_3\sum x_i^3 + \ldots] \qquad \ldots(15\cdot61a)$$

Thus both the methods yield identical estimators if *MLE*'s are obtained as linear functions of the moments.

Example 15·43. *Estimate α and β in the case of Pearson's Type III distribution by the method of moments.*

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ 0 \le x < \infty$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1988*]

Solution. We have

$$\mu_r' = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^r x^{\alpha-1} e^{-\beta x}\,dx = \frac{\beta^\alpha}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+r)}{\beta^{\alpha+r}} = \frac{\Gamma(\alpha+r)}{\Gamma(\alpha)\,\beta^r}$$

$$\mu_1' = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha).\beta} = \frac{\alpha}{\beta},\quad \mu_2' = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\,\beta^2} = \frac{(\alpha+1)\,\alpha}{\beta^2}$$

$$\frac{\mu_2'}{\mu_1'^2} = \frac{\alpha+1}{\alpha} = \frac{1}{\alpha} + 1$$

$$\Rightarrow \qquad \alpha = \frac{\mu_1'^2}{\mu_2' - \mu_1'^2}, \quad \beta = \frac{\alpha}{\mu_1'} = \frac{\mu_1}{\mu_2' - \mu_1'^2}$$

$$\text{Hence} \qquad \hat{\alpha} = \frac{m_1'^2}{m_2' - m_1'^2} \quad \text{and} \quad \hat{\beta} = \frac{m_1'}{m_2' - m_1'^2}$$

where m_1' and m_2' are the sample moments.

Example 15·44. *For the double Poisson distribution :*

$$p(x) = P(X = x) = \frac{1}{2} \cdot \frac{e^{-m_1} m_1^x}{x\,!} + \frac{1}{2} \cdot \frac{e^{-m_2} m_2^x}{x\,!}\,;\; x = 0, 1, 2, \ldots$$

show that the estimates for m_1 and m_2 by the method of moments are :

$$\mu_1' \pm \sqrt{\mu_2' - \mu_1' - \mu_1'^2}$$

[***Delhi Univ. B.Sc. (Stat. Hons.), 1993***]

Solution. We have

$$\mu_1' = \sum_{x=0}^{\infty} x.p(x) = \frac{1}{2} \sum_{x=0}^{\infty} x. \frac{e^{-m_1} m_1^x}{x\,!} + \frac{1}{2} \sum_{x=0}^{\infty} x \cdot \frac{e^{-m_2}. m_2^x}{x\,!}$$

$$= \frac{1}{2} m_1 + \frac{1}{2} m_2 \qquad \ldots(*)$$

(since the first and second summations are the means of Poisson distributions with parameters m_1 and m_2 respectively).

$$\mu_2' = \sum_{x=0}^{\infty} x^2 . p(x)$$

$$= \frac{1}{2}\left[\sum_{x=0}^{\infty} x^2 \cdot \left(\frac{e^{-m_1} m_1^x}{x\,!}\right) + \sum_{x=0}^{\infty} x^2 \cdot \left(\frac{e^{-m_2} m_2^x}{x\,!}\right)\right]$$

$$= \frac{1}{2}\left[(m_1^2 + m_1) + (m_2^2 + m_2)\right] \qquad [c.f. \; \S\, 7{\cdot}3{\cdot}3]$$

$$\mu_2' = \frac{1}{2}\left[(m_1 + m_2) + (m_1^2 + m_2^2)\right] \qquad \ldots(**)$$

$$= \frac{1}{2}\left[2\mu_1' + m_1^2 + (2\mu_1' - m_1)^2\right] \qquad [\text{Using } (*)]$$

$$= \frac{1}{2}\left[2\mu_1' + m_1^2 + 4\mu_1'^2 + m_1^2 - 4m_1\,\mu_1'\right]$$

$$\mu_2' = \mu_1' + m_1^2 + 2\mu_1'^2 - 2\mu_1' m_1 \;\Rightarrow\; m_1^2 - 2m_1\mu_1' + (2\mu_1'^2 + \mu_1' - \mu_2') = 0$$

$$\Rightarrow \hat{m}_1 = \frac{2\mu_1' \pm \sqrt{4\mu_1'^2 - 4(2\mu_1'^2 + \mu_1' - \mu_2')}}{2} = \mu_1' \pm \sqrt{\mu_2' - \mu_1' - \mu_1'^2}$$

Similarly on substituting for m_1 in terms of m_2 from (*) in (**), we get

$$m_2^2 - 2m_2\mu_1' + (2\mu_1'^2 + \mu_1' - \mu_2') = 0$$

Solving for m_2, we will get

$$\hat{m}_2 = \mu_1' \pm \sqrt{\mu_2' - \mu_1' - \mu_1'^2}$$

Example 15·45. *A random variable X takes the values 0, 1, 2, with respective probabilities*

$$\frac{\theta}{4N}+\frac{1}{2}\left(1-\frac{\theta}{N}\right),\ \frac{\theta}{2N}+\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) \quad \text{and} \quad \frac{\theta}{4N}+\frac{1-\alpha}{2}\left(1-\frac{\theta}{N}\right)$$

where N is a known number and α, θ are unknown parameters. If 75 independent observations on X yielded the values 0, 1, 2 with frequencies 27, 38, 10 respectively, estimate θ and α by the method of moments.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

Solution.

$$E(X)=0\cdot\left[\frac{\theta}{4N}+\frac{1}{2}\left(1-\frac{\theta}{N}\right)\right]+1\cdot\left[\frac{\theta}{N}+\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)\right]$$

$$+2\left[\frac{\theta}{4N}+\frac{1-\alpha}{2}\left(1-\frac{\theta}{N}\right)\right]$$

$$=\frac{\theta}{N}+\left(1-\frac{\theta}{N}\right)\left[\frac{\alpha}{2}+(1-\alpha)\right]$$

$$\Rightarrow \quad \mu_1'=\frac{\theta}{N}+\left(1-\frac{\theta}{N}\right)\left(1-\frac{\alpha}{2}\right)$$

$$=1-\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right) \qquad \ldots(*)$$

$$E(X^2)=1^2\cdot\left[\frac{\theta}{2N}+\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)\right]+2^2\cdot\left[\frac{\theta}{4N}+\frac{1-\alpha}{2}\left(1-\frac{\theta}{N}\right)\right]$$

$$=\frac{3\theta}{2N}+\left(1-\frac{\theta}{N}\right)\left[\frac{\alpha}{2}+2(1-\alpha)\right]$$

$$=\frac{3\theta}{2N}+\left(1-\frac{\theta}{N}\right)\left(2-\frac{3\alpha}{2}\right)$$

$$\Rightarrow \quad \mu_2'=2-\frac{\theta}{2N}-\frac{3}{2}\alpha\left(1-\frac{\theta}{N}\right) \qquad \ldots(**)$$

The sample frequency distribution is :

x	0	1	2
f	21	38	10

$$\mu_1'=\frac{1}{N}\Sigma fx=\frac{1}{75}(38+20)=\frac{58}{75}$$

$$\mu_2'=\frac{1}{N}\Sigma fx^2=\frac{1}{75}(38+40)=\frac{78}{75}$$

Equating the sample moments to theoretical moments, we get

$$1-\frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)=\frac{58}{75}$$

$$\Rightarrow \quad \frac{\alpha}{2}\left(1-\frac{\theta}{N}\right)=1-\frac{58}{75}=\frac{17}{75} \qquad \ldots(***)$$

Substituting in (**), we get

$$2-\frac{\theta}{2N}-3\times\frac{17}{75}=\frac{78}{75} \quad\Rightarrow\quad \hat{\theta}=\frac{42}{75}N$$

Substituting in (***), we get

$$\frac{\alpha}{2}\left(1-\frac{42}{75}\right)=\frac{17}{75} \quad\Rightarrow\quad \hat{\alpha}=\frac{34}{33}$$

15·14. Method of Least Squares.* The principle of least squares is used to fit a curve of the form

$$y=f(x, a_0, a_1, \ldots, a_n) \qquad \ldots(15\cdot62)$$

where a_i's are unknown parameters, to a set of n sample observations (x_i, y_i); $i = 1, 2, \ldots, n$ from a bivariate population. It consists in minimising the sum of squares of residuals, *viz.*,

$$E=\sum_{i=1}^{n}\left[y_i-f(x_i, a_0, a_1, \ldots, a_n)\right]^2 \qquad \ldots(15\cdot63)$$

subject to variations in $a_0, a_1, \ldots, a_n$.

The normal equations for estimating $a_0, a_1, \ldots, a_n$ are given by

$$\frac{\partial E}{\partial a_i}=0;\ i=1, 2, \ldots, n \qquad \ldots(15\cdot64)$$

Remarks. 1. In chapter 9, we have discussed in detail the method of least squares for fitting linear regression (§ 9·1·1), polynomial regression (§ 9·1·3) and the exponential family of curves reducible to linear regression (§ 9·3). In chapter 10 § 10·12·1, we have discussed the method of fitting multiple linear regression.

2. If we are estimating $f(x, a_0, a_1, \ldots, a_n)$ as a linear function of the parameters $a_0, a_1, \ldots, a_n$, the x's being known given values, the least square estimators obtained as linear functions of the y's will be MVU estimators.

EXERCISE 15(b)

1. (*a*) State and explain the principle of maximum likelihood for estimation of population parameter.

(*b*) (*i*) Describe the M.L. method of estimation and discuss five of its optimal properties.

(*ii*) Examine a situation when M.L. method fails and explain how you tackle such situations.

(*c*) Define the likelihood function for a random sample drawn from (*i*) a discrete population, (*ii*) a continuous population.

Find the likelihood function for a random sample of size n from each of the following populations :

(*a*) Normal (m, σ^2), (*b*) Binomial (n, p), (*c*) Poisson (μ), (*d*) Uniform on (a, b). [*Calcutta Univ. B.Sc. (Maths. Hons.), 1991*]

---

* For detailed discussion see Chapter 9.

2. (*a*) A random variable X takes the values 0 and 1 with respective probabilities p and $1-p$. Obtain on the basis of random sample of size n, the maximum likelihood estimator of p.

(*b*) Obtain the maximum likelihood estimator for the distribution having the probability mass function :

$$f(x, \theta) = \theta^x (1-\theta)^{1-x}, x = 0, 1 ; 0 \le \theta \le 1$$

[***Calcutta Univ. B.Sc. (Maths. Hons.), 1986***]

(*c*) Obtain the maximum likelihood estimator of θ in the following cases :

(*i*) $f(x, \theta) = \dfrac{1}{\theta} . \exp(-x/\theta) ; x \ge 0, \theta > 0$

(*ii*) $f(x, \theta) = {}^nC_x \, \theta^x (1-\theta)^{n-x} ; x = 0, 1, 2, ..., n$

3. Suppose that X has a distribution $N(\mu, \sigma^2)$, that is, the p.d.f of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Using M.L. estimation, determine μ and σ^2. What conclusions do you draw on the nature of the result so obtained ?

4. (*a*) Explain the technique of the method of maximum likelihod and give a formula for the large sample standard error of the maximum-likelihood estimator.

(*b*) For the distribution with p.d.f.

$f(x, \theta) = \theta e^{-\theta x}, (x \ge 0 ; \theta > 0)$, find the maximum likelihood estimators of θ and $E(X)$, and obtain their large-sample standard errors.

(*c*) X is a random variable such that

$$\begin{aligned} P(X \le x) &= 0, \quad \text{for } x < 0 \\ &= 1 - e^{-x\theta}, \quad \text{for } x \ge 0 \end{aligned}$$

Based on n independent observations on X, obtain the maximum likelihood estimator of $E(X)$.

5. (*a*) Let $X_1, X_2, ..., X_n$ be a random sample from the distribution with probability density function :

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}; 0 < x < \infty, \; 0 < \theta < \infty$$

Find the maximum likelihood estimator of θ.

[***Madras Univ. B.Sc. Sept., 1988***]

(*b*) For the distribution :

$$dF(x) = \frac{1}{\theta^p \Gamma(p)} \exp(-x/\theta)\, x^{p-1} ; 0 \le x < \infty, p > 0, \theta > 0$$

where p is known, find out the maximum likelihood estimate of θ on the basis of a random sample of size n from the distribution. Find the variance of the estimate.

6. (*a*) If x_i $(i = 1, 2, ... n)$ is an observed random sample from the distribution having p.d.f.

$$f_\lambda(x) = \frac{\lambda^{k+1} x^k \exp(-\lambda x)}{\Gamma(k+1)}, x > 0$$

where $\lambda > 0$ and k is a known constant, show that the ML estimator $\hat{\lambda}$ for λ is $(k+1)/\bar{x}$ Show that the corresponding estimator is biased but consistent and that its asymptotic distribution for large n is

$$N\left(\lambda, \lambda^2/[n(k+1)]\right).$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1986*]

(*b*) Derive the *MLE* of the mean $\dfrac{\alpha}{\alpha+2}$ of the beta distribution :

$$f(x) = [B(\alpha, 2)]^{-1}\, x^{\alpha-1}\,(1-x),\; 0 < x \leq 1,\; \alpha > 0.$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

7. (*a*) From a sample of size n from the population of X, determine the maximum likelihood estimates of the parameters a and b of the probability density

$$f(x) = \text{Constant} \exp\,[-(x-a)/b];\; x \geq a,\; b > 0,\; -\infty < a < \infty$$

[*Calcutta Univ. B.Sc. (Maths Hons.), 1991*]

(*b*) Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution with p.d.f.

$$f(x;\theta_1,\theta_2) = \begin{cases} \dfrac{1}{\theta_2} e^{-(x-\theta_1)/\theta_2}, & x \geq \theta_1,\; -\infty < \theta_1 < \infty,\; \theta_2 > 0 \\ 0, \text{ elsehwere} & \end{cases}$$

Obtain the maximum likelihood estimators for θ_1 and θ_2.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992*]

(*c*) Given a sample of n independent observations from the distribution with density : $\quad f(x, \theta_1, \theta_2) = \theta_2^{-1} \exp\left[-(x-\theta_1)/\theta_2\right],\; \theta_1 \leq x < \infty$

Find the maximum-likelihood estimator of θ_2 when θ_1 is known and the maximum likelihood estimator of θ_1 when θ_2 is known and also the joint maximum likelihood estimators of θ_1 and θ_2. Comment on the estimators you obtain.

8. (*a*) A random variable X has the probability density function :

$$f(x) = (\beta+1)\,x^{\beta}, \text{ for } (0 < x < 1), (\beta > -1).$$
$$= 0, \qquad \text{otherwise.}$$

Based on n-independent observations on X, obtain the maximum likelihood estimator of β and an unbiased estimator of $(\beta+1)/(\beta+2)$, when $\beta \neq -2$.

(*b*) A random variable X has a distribution with density function

$$f(x) = (\alpha+1)\,x^{\alpha},\; (0 \leq x \leq 1,\; \alpha > -1)$$
$$= 0, \qquad \text{otherwise}$$

and a random sample of size 8 produces the data :

0·2, 0·4, 0·8, 0·5, 0·7, 0·9, 0·8, 0·9.

Find the maximum likelihood estimate of the unknown parameter α, it being given that ln (0·0145152 $\simeq -4{\cdot}2326$ (ln denotes natural logarithm).

[*Burdwan Univ. B.Sc. (Hons.), 1989*]

(c) Find the MLE of θ for a random sample of size n from the distribution :

$$f(x, \theta) = (\theta + 1)\, x^{\theta}, \quad 0 \le x \le 1$$
$$= 0, \qquad \text{otherwise}$$

Show that it is also sufficient statistic for θ.

Ans. $$MLE\,(\hat{\theta}) = \left[-\frac{n}{\sum_{i=1}^{n} \log x_i} - 1 \right] \qquad \ldots(*)$$

$$T = \prod_{i=1}^{n} x_i, \text{ is sufficient estimator for } \theta$$

$\Rightarrow$ $\hat{\theta} = \left[\dfrac{-n}{\log (\prod_i x_i)} - 1 \right]$, being a one to one function of sufficient statistic, is also a sufficient statistic for θ.

9. (a) Obtain the MLE for the parameter θ in a random sample of size n from the uniform population $U[0, \theta]$.

Ans. $\hat{\theta} = x_{(n)}$, the largest sample observation.

(b) Show by means of an example, that MLE are not, in general unique.

Ans. See Example 15·34.

(c) Show that in a random sample from a distribution with p.d.f.

$$f(x, \theta) = \theta e^{-\theta x}, x \ge 0$$

$1/\bar{X}$ is the MLE for θ and has greater variance than the unbiased estimator $(n - 1)/(n\bar{X})$.

Hint. MLE $\hat{\theta} = \dfrac{1}{\bar{X}} = \dfrac{n}{T}$, $T = \sum_{i=1}^{n} X_i \Rightarrow n\bar{X} = T$

$$X_i, (i = 1, 2, \ldots, n) \text{ are } i.i.d.\ \gamma(\theta, 1)$$

$$\Rightarrow \quad T = \sum_i X_i \sim \gamma(\theta, n)$$

$$E\left[\frac{n-1}{n\bar{X}}\right] = E\left[\frac{n-1}{T}\right] = (n-1)\, E(1/T) = \theta.$$

$$\text{Var}\left(\frac{n-1}{n\bar{X}}\right) = \left(\frac{n-1}{n}\right)^2 \text{Var}\left(\frac{1}{\bar{X}}\right) < \text{Var}\left(\frac{1}{\bar{X}}\right) = \text{Var}\,\hat{\theta}$$

10. (a) Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with density :

$$f(x, \theta) = \frac{1}{2} \exp\left[-|x - \theta| \right], -\infty < x < \infty.$$

Find the estimator for θ based on the method of maximum likelihood.

[*Madras Univ. B.Sc., 1989*]

Hint. $L = \left(\dfrac{1}{2}\right)^n \exp\left[-\sum_{i=1}^{n} |x_i - \theta| \right]$ is maximum, if $\sum_{i=1}^{n} |x_i - \theta|$ is minimum. $\Rightarrow \hat{\theta} =$ Median of $(x_1, x_2, \ldots, x_n)$.

(*b*) Obtain the maximum likelihood estimator of θ based on a random sample of size n from the population with p.d.f.

(*i*) $f(x, \theta) = e^{-(x-\theta)}$; $\theta \le x < \infty, -\infty < \theta < \infty$

(*ii*) $f(x, \theta) = \theta x^{\theta-1}$; $0 < x < 1, 0 < \theta < \infty$.

Examine in each case, whether θ is unbiased.

Hint. (*i*) L is maximum if $\sum_{i=1}^{n} (x_i - \theta)$ is minimum.

$\Rightarrow$ Each deviation $(x_i - \theta), i = 1, 2, \ldots, n$ is minimum $\Rightarrow$ $\hat{\theta} = x_{(1)}$.

11. (*a*) Explain what is meant by an estimate of a population parameter. Find the maximum likelihood estimate of the parameter θ of a population having density function :

$$2(\theta - x)/\theta^2, (0 < x < \theta)$$

for a sample of unit size and examine whether the estimate so obtained is biased or not. [*Calcutta Univ. B.Sc. (Maths. Hons.), 1987*]

Ans. $\hat{\theta} = 2x$; biased.

(*b*) Obtain Maximum Likelihood Estimator of θ for the distribution :

$$f(x, \theta) = \frac{C_o \theta^x}{x!} ; x = 0, 1, 2, \ldots; \theta > 0,$$

C_o is a constant. Also write the Maximum Likelihood Estimator of $3\theta^2 + 4\theta + 5$. (*Agra Univ. B.Sc., 1988*)

Hint. For MLE of $3\theta^2 + 4\theta + 5$, use Invariance Property of MLE (*c.f.* Theorem 15·17)

(*c*) A population has a density function given by :

$$f(x) = 2v\sqrt{\frac{v}{\pi}}\, x^2 e^{-vx^2} ; -\infty < x < \infty$$

Find the maximum likelihood estimate for v.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1988*]

12. (*a*) Consider a population made up of 3 different types of individuals occurring in the population with probabilities θ^2, $2\theta(1-\theta)$ and $(1-\theta)^2$, respectively where $0 < \theta < 1$. Let n_1, n_2 and n_3 denote the respective random sample sizes of the above three types of individuals. Determine the maximum likelihood estimator for θ. [*Rajasthan PCS, 1989*]

(*b*) Obtain the maximum likelihood estimate of θ, if the variable takes the values 1, 2, 3 and 4 with probabilities $(1-\theta)/2$, $(1-\theta)/2$, $\theta(1-\theta)$ and θ^2 respectively and the observed frequencies are n_1, n_2, n_3 and n_4 respectively.

13. In life-testing it is sometimes assumed that the life-time of an item is a random variable which is greater than or equal to x with probability

$$\exp\left[-\left(\frac{x}{\theta}\right)^m\right],$$

$x \ge 0, m > 0$ is known and $\theta > 0$ is unknown. Suppose n such items are tested and yield $X_1, X_2, \ldots, X_n$ as their times of "death".

Find the maximum likelihood estimate of θ.

14. X_1, X_2, X_3, X_4 are independent normal random variables with means $\alpha + \beta, \alpha - \beta, \alpha + 2\beta, \alpha - \beta$ respectively and a common variance σ^2, on the basis of one observation on each X_i ; obtain the maximum likelihood estimators of α, β and σ^2. What is the asymptotic variance of α^2 ?

[*Bharatiyan Univ. M.Sc. (Maths), 1993*]

15. (*a*) For the bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ find the maximum likelihood estimators

(*i*) of σ_1^2, σ_2^2 and ρ when μ_1 and μ_2 are known,

(*ii*) of all five parameters of the distribution.

(*b*) Describe clearly the important properties to be possessed by a good estimator.

If (x_i, y_i), $(i = 1, 2, \ldots, n)$ come from a bivariate normal population with zero means, unit variances and co-efficient of correlation ρ, obtain the maximum likelihood estimator of ρ.

16. (*a*) Show that the most general continuous distribution for which the M.L.E. of a parameter θ is the sample harmonic mean is :

$$f(x, \theta) = \exp\left[\frac{1}{x}\left\{\theta\frac{\partial \psi}{\partial \theta} - \psi(\theta)\right\} - \frac{\partial \psi}{\partial \theta} + \xi(x)\right]$$

where $\psi(\theta)$ and $\xi(x)$ are arbitrary functions of θ and x respectively.

(*b*) Explain the principle of maximum likelihood estimation. Give examples to show that MLE need not be unique and also not necessarily unbiased.

Show that the most general form of the distribution for which the sample arithmetic mean $\overline{X}$ is the MLE of θ has the p.d.f.

$$f(x, \theta) = \exp\left[(x - \theta)A'(\theta) + A(\theta) + B(x)\right]$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

17. (*a*) Suppose that distribution of X is represented by the function :

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x\,!}\,;\; x = 0, 1, 2, \ldots$$

where $\lambda > 0$. Given a random sample of size n, show that the sample mean is the maximum likelihood estimate of λ. Show further that this estimate is (*i*) best unbiased, and (*ii*) consistent. [*Delhi Univ. M.A. (Eco.), 1986*]

(*b*) Consider the estimation of the Poisson parameter from a random sample.

(*i*) Work out the maximum likelihood estimator and its variance.

(*ii*) Work out the Cramer – Rao Lower bound and show that it is equal to the variance worked out in (*i*). Comment on the significance of this result.

[*Delhi Univ. M.A. (Eco.), 1990*]

18. X is a discrete random variable and

$$P(X = r) = (1 - p)\,p^{r-1};\; r = 1, 2, 3, \ldots$$

Find the MLE of p based on a random sample of n observations and its variance in large samples.

Show that the variance attains the lower bound of C.R. inequality.

19. Explain the terms : (*i*) sufficient estimator, (*ii*) efficient estimator, (*iii*) Cramer-Rao lower bound to the variance of an estimator, (*iv*) maximum likelihood estimator; and describe the relations amongst these four concepts.

20. (*a*) Describe the method of moments for estimating the parameters. What are the properties of the estimates obtained by this method ?

(*b*) Let $(X_1, X_2, \ldots, X_n)$ be a random sample from the p.d.f.

$$\begin{aligned} f(x, \theta) &= \theta e^{-\theta x}, \; 0 < x < \infty, \; \theta > 0; \\ &= 0, \quad \text{elsewhere} \end{aligned}$$

Estimate θ using the method of moments.

(*Madras Univ. B.Sc., 1988*)

21. $X_1, X_2, \ldots, X_n$ is a random sample from

$$\begin{aligned} f(x\,;a, b) &= \frac{1}{b-a}\,; \quad a < x < b \\ &= 0, \quad \text{elsewhere} \end{aligned}$$

Find estimates of a and b by the method of moments.

[*Gujarat Univ. B.Sc. Oct., 1993*]

22. Explain the methods of estimation-method of moments and maximum likelihood. Do these lead to the same estimates in respect of the standard deviation of a normal population ? Examine the properties of the estimates from the point of view of consistency and unbiasedness.

23. (*a*) Estimate θ in the density function

$$f(x, \theta) = (1 + \theta)\, x^{\theta}\,; \; 0 < x < 1$$

by the method of moments and obtain the standard error of the estimator.

(*b*) The sample values from population with p.d.f.

$$f(x) = (1 + \theta)\, x^{\theta}, \; 0 < x < 1, \; \theta > 0,$$

are given below :

0·46, 0·38, 0·61, 0·82, 0·59, 0·53, 0·72, 0·44, 0·59, 0·60

Find the estimate of θ by (*i*) method of moments and (*ii*) maximum likelihood estimation.

24. (*a*) For the distribution with probability function :

$$f(x, \theta) = \frac{e^{-\theta}\,\theta^{x}}{x\,!\,(1 - e^{-\theta})}\,; \; x = 1, 2, 3, \ldots$$

obtain the estimate of θ by the method of moments.

(*b*) For the following probability function :

$$f(x, p) = \binom{3}{x} \frac{p^{x}\,(1-p)^{3-x}}{1 - (1-p)^{3}}, \; [x = 1, 2, 3].$$

obtain the estimator of p by the method of moments, if the frequencies at $x = 1, 2$ and 3 are respectively 22, 20 and 18.

25. Let $x_1, x_2, \ldots, x_n$ be a sample from a distribution with density function :

$$f_{\theta}(x) = \theta(\theta + 1)\, x^{\theta - 1}\,(1 - x), \; 0 < x < 1, \; \theta > 0$$

Determine the estimate of θ by the method of moments.

[*Indian Civil Services, 1981*]

26. Explain the method of minimum chi-square in estimation, with a suitable example. [*Madras Univ. B.Sc., March 1989*]

27. Describe the method of moments and discuss when the estimates obtained by the method of moments are identical with those of maximum likelihood estimates.

Estimate α and β by the method of moments for the distribution :

$$f(x;\alpha,\ \beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}\,e^{-\beta x}\ ,\ 0\leq x<\infty.$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1983*]

28. State the conditions under which Maximum Likelihood Estimators of the parameters are identical with those given by the method of moments.

Examine if the MLEs of the parameter(s) are identical with those obtained by the method of moments in random sampling from the following distributions :

(*i*) $f(x,\theta)=\frac{1}{\theta}\cdot\exp\left(-\frac{x}{\theta}\right)\ ;\ 0<x<\infty$

(*ii*) $f(x,\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-(x-\mu)^2/2\sigma^2\right]\ ;\ -\infty<x<\infty.$

Ans. (*i*) *MLE* $(\hat{\theta})=\bar{x}=\hat{\theta}$ (Method of Moments)

(*ii*) MLE $(\hat{\mu})=\bar{X}=\hat{\mu}$ (Method of Moments)

MLE $(\hat{\sigma^2})=s^2$ (sample variance) $=\hat{\sigma}^2$ (Method of Moments).

29. Independent samples of sizes n_1 and n_2 are taken from two normal populations with equal means μ and variances respectively equal to $\lambda\sigma^2$, σ^2. Find the maximum likelihood estimator of μ based on (n_1+n_2) sample observations and show that its large sample variance is

$$Var\,(\hat{\mu})=\sigma^2\Big/\left(\frac{n_1}{\lambda}+n_2\right)$$

Hence show that the unbiased estimator, $t=(n_1\bar{x}_1+n_2\bar{x})/(n_1+n_2)$ has efficiency, $\dfrac{\lambda(n_1+n_2)^2}{(n_1\lambda+n_2)(n_1+n_2\lambda)}$ which attains the value 1 if and only if $\lambda=1$.

Ans. MLE $(\hat{\mu})=\left(\frac{n_1\bar{x}_1}{\lambda}+n_2\bar{x}_2\right)\Big/\left(\frac{n_1}{\lambda}+n_2\right)$

OBJECTIVE TYPE QUESTIONS

1. Comment on the following statements :

(*i*) In case of the Poisson distribution with parameter λ, $\bar{x}$ is sufficient for λ.

(*ii*) If $(X_1, X_2, \ldots, X_n)$ be a sample of independent observations from the uniform distribution on $(\theta, \theta + 1)$, then the maximum likelihood estimator of θ is unique.

(*iii*) A maximum likelihood estimator is always unbiased.

(*iv*) Unbiased estimator is necessarily consistent

(*v*) A consistent estimator is also unbiased.

(*vi*) An unbiased estimator whose variance tends to zero as sample size increases is consistent.

(*vii*) If t is a sufficient statistic for θ then $f(t)$ is a sufficient statistic for $f(\theta)$.

(*viii*) If t_1 and t_2 are two independent estimators of θ, then $t_1 + t_2$ is less efficient then both t_1 and t_2.

(*ix*) If T is consistent estimator of a parameter θ, then $aT + b$ is a consistent estimator of $a\theta + b$, where a and b are constants.

(*x*) If x is the number of successes in n independent trials with a constant probability p of success in each trial, then x/n is a consistent estimator of p.

II. Fill in the blanks :

(*i*) In a random sample of size n from a population with mean μ, the sample mean ($\bar{x}$) is ... estimate of ...

(*ii*) The sample median is ... estimate for the mean of normal population.

(*iii*) An estimator $\hat{\theta}$ of a parameter θ is said to be unbiased if ...

(*iv*) The variance s^2 of a sample of size n is a ... estimator of population variance σ^2.

(*v*) If a sufficient estimator exists, it is a function of the ... estimator.

(*vi*) ... estimate may not be unique.

III. (*a*) Give example of a statistic t which is unbiased for a parameter θ but t^2 is not unbiased for θ^2.

(*b*) Give example of an M.L. estimator which is not unbiased.

IV. What is the relationship between a sufficient estimator and a maximum likelihood estimator ?

V. (*i*) If $\bar{x}$ is an unbiased estimator for the population mean μ, state which of the following are unbiased estimators for μ^2 :

$$(a)\ \bar{x}^2,\quad (b)\ \bar{x}^2 - \frac{\sigma^2}{n}\ (\sigma^2 \text{ is known/unknown}).$$

(*ii*) If t is the maximum likelihood estimator for θ, state the condition under which $f(t)$ will be the maximum likelihood estimator for $f(\theta)$.

(*iii*) Write down the condition for the Cramer-Rao lower bound for the variance of an unbiased estimator to be attained.

(*iv*) Write down the general form of the distribution admitting sufficient statistic.

VI. A random variable X takes the values 1, 2, 3 and 4, each with probability $\frac{1}{4}$. A random sample of three values of x is taken, $\bar{x}$ is the mean and m is the median of this sample. Show that both $\bar{x}$ and m are unbiased estimators of the mean of the population, but $\bar{x}$ is more efficient than m. Compare their efficiencies.

VII. Give an example of estimates which are

(*i*) Unbiased and efficient, (*ii*) Unbiased and inefficient.

15·15. Confidence Interval and Confidence Limits. Let x_i, $(i = 1, 2, \ldots, n)$ be a random sample of n observations from a population involving a single unknown parameter θ (say). Let $f(x, \theta)$ be the probability function of the parent distribution from which the sample is drawn and let us suppose that this distribution is continuous. Let $t = t(x_1, x_2, \ldots, x_n)$, a function of the sample values be an estimate of the population parameter θ, with the sampling distribution given by $g(t, \theta)$.

Having obtained the value of the statistic t from a given sample, the problem is, "Can we make some reasonable probability statements about the unknown parameter θ in the population, from which the sample has been drawn ?" This question is very well answered by the technique of *Confidence Interval* due to Neyman and is obtained below :

We choose once for all some small value of α (5% or 1%) and then determine two constants say, c_1 and c_2 such that

$$P(c_1 < \theta < c_2 \mid t) = 1 - \alpha \qquad \ldots(15{\cdot}65)$$

The quantities c_1 and c_2, so determined, are known as the *confidence limits* or *fiducial limits* and the interval $[c_1, c_2]$ within which the unknown value of the population parameter is expected to lie, is called the *confidence interval* and $(1 - \alpha)$ is called the *confidence coefficient*.

Thus if we take $\alpha = 0{\cdot}05$ (or $0{\cdot}01$), we shall get 95% (or 99%) confidence limits.

How to find c_1 and c_2 ? Let T_1 and T_2 be two statistics such that

$$P(T_1 > \theta) = \alpha_1 \qquad \ldots(15{\cdot}66)$$

and

$$P(T_2 < \theta) = \alpha_2 \qquad \ldots(15{\cdot}66a)$$

where α_1 and α_2 are constants independent of θ. (15·66) and (15·66*a*) can be combined to give

$$P(T_1 < \theta < T_2) = 1 - \alpha, \qquad \ldots(15{\cdot}66b)$$

where $\alpha = \alpha_1 + \alpha_2$. Statistics T_1 and T_2 defined in (15·66) and (15·66*a*) may be taken as c_1 and c_2 defined in (15·65).

For example, if we take a large sample from a normal population with mean μ and standard deviation σ, then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

and

$$P(-1{\cdot}96 < Z < 1{\cdot}96) = 0{\cdot}95,$$

[From Normal Probability Tables]

$$\Rightarrow \qquad P\left(-1{\cdot}96 < \frac{\bar{x} - \mu}{\sigma\sqrt{n}} < 1{\cdot}96\right) = 0{\cdot}95$$

$$\Rightarrow \quad P\left[\bar{x} - 1{\cdot}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{\cdot}96\frac{\sigma}{\sqrt{n}}\right] = 0{\cdot}95$$

Thus $\bar{x} \pm 1{\cdot}96 \dfrac{\sigma}{\sqrt{n}}$ are 95% confidence limits for the unknown parameter μ, the population mean and the interval

$\left[\bar{x} - 1{\cdot}96\dfrac{\sigma}{\sqrt{n}}, \bar{x} + 1{\cdot}96\dfrac{\sigma}{\sqrt{n}}\right]$ is called the 95% confidence interval.

Also $$P(-2{\cdot}58 < Z < 2{\cdot}58) = 0{\cdot}99$$

$$\Rightarrow \quad P\left(-2{\cdot}58 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 2{\cdot}58\right) = 0{\cdot}99$$

$$\Rightarrow \quad P\left(\bar{x} - 2{\cdot}58\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2{\cdot}58\frac{\sigma}{\sqrt{n}}\right) = 0{\cdot}99$$

Hence 99% confidence limits for μ are $\bar{x} \pm 2{\cdot}58 \dfrac{\sigma}{\sqrt{n}}$ and 99% confidence interval for μ is $\left[\bar{x} - 2{\cdot}58\dfrac{\sigma}{\sqrt{n}}, \bar{x} + 2{\cdot}58\dfrac{\sigma}{\sqrt{n}}\right]$.

Remarks 1. Usually σ^2 is not known and its unbiased estimate S^2 obtained from the samples, is used. However if n is *small*,

$$Z = \frac{\bar{x} - \mu}{S/\sqrt{n}} \text{ is not } N(0, 1)$$

and in this case the confidence limits and confidence intervals for μ are obtained by using Student's 't' distribution.

2. It can be seen that in many cases there exist more than one set of confidence intervals with the same confidence coefficient. Then the problem arises as to which particular set is to be regarded as better than the others in some useful sense and in such cases we look for the shortest of all the intervals.

Example 15·45. *Obtain 100 (1 − α)% confidence intervals for the parameters (a) θ and (b) σ², of the normal distribution*

$$f(x, \theta; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\left(-\frac{1}{2}\left(\frac{x - \theta}{\sigma}\right)^2\right)\right], -\infty < x < \infty$$

Solution. Let X_i, $(i = 1, 2, \ldots, n)$ be a random sample of size n from the density $f(x\,;\theta, \sigma)$ and let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

(*a*) The statistic : $$t = \frac{\bar{X} - \theta}{S/\sqrt{n}}$$

follows student's t-distribution with $(n - 1)$ degrees of freedom. Hence $100(1 - \alpha)\%$ confidence limits for θ are given by

$$P[|t| \leq t_\alpha] = 1 - \alpha$$

$$\Rightarrow \qquad P\left[|\bar{X} - \theta| \leq \frac{S}{\sqrt{n}} t_\alpha\right] = 1 - \alpha$$

$$\Rightarrow \quad P\left[\bar{X} - t_\alpha \cdot \frac{S}{\sqrt{n}} \leq \theta \leq \bar{X} + t_\alpha \cdot \frac{S}{\sqrt{n}}\right] = 1 - \alpha \qquad \ldots(567)$$

where t_α is the tabulated value of t for $(n - 1)$ d.f. at significance level 'α'. Hence the required confidence interval for θ is :

$$\left(\bar{X} - t_\alpha \frac{S}{\sqrt{n}}, \bar{X} + t_\alpha \frac{S}{\sqrt{n}}\right)$$

(*b*) *Case (i) θ is known and equal to μ (say).*

Then $$\frac{\Sigma(X_i - \mu)^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n)}$$

If we define χ_α^2 as the value of χ^2 such that

$$P(\chi^2 > \chi_\alpha^2) = \int_{\chi_\alpha^2}^{\infty} p(\chi^2)\, d\chi^2 = \alpha \qquad \ldots(*)$$

where $p(\chi^2)$ is the p.d.f. of χ^2-distribution with n d.f., then the required confidence interval is given by

$$P[\chi^2_{1-(\alpha/2)} \leq \chi^2 \leq \chi^2_{\alpha/2}] = 1 - \alpha$$

$$\Rightarrow \qquad P\left[\chi^2_{1-(\alpha/2)} \leq \frac{ns^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right] = 1 - \alpha \qquad \ldots(**)$$

Now $$\frac{ns^2}{\sigma^2} \leq \chi^2_{\alpha/2} \Rightarrow \frac{ns^2}{\chi^2_{\alpha/2}} \leq \sigma^2$$

and $$\chi^2_{1-(\alpha/2)} \leq \frac{ns^2}{\sigma^2} \Rightarrow \sigma^2 \leq \frac{ns^2}{\chi^2_{1-(\alpha/2)}}.$$

Hence (**) gives

$$P\left[\frac{ns^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{1-(\alpha/2)}}\right] = 1 - \alpha \qquad \ldots(***)$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-(\alpha/2)}$ are obtained from (*) by using n d.f.

Thus *e.g.*, 95% confidence interval for σ^2 is given by

$$P\left[\frac{ns^2}{\chi^2_{0.025}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{0.975}}\right] = 0.95$$

Case (ii). θ is unknown. In this case the statistic

$$\frac{\Sigma(X_i - \bar{X})^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

Here also confidence interval for σ^2 is given by (***) where now χ^2_α is the significant value of χ^2 [as defined in (*)] for $(n - 1)$ d.f. at the significance level 'α'.

Example 15·46. *Show tnat the largest observations L of a sample of n observations from a rectangular distribution with density function :*

$$f(x, \theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta \qquad \ldots(*)$$
$$= 0, \text{ otherwise}$$

has the distribution

$$dG(L) = n\left(\frac{L}{\theta}\right)^{n-1} \cdot \frac{dL}{\theta}, \; 0 \le L \le \theta$$

Show that the distribution of $V = L/\theta$ is given by p.d.f.

$$h(v) = nv^{n-1}, 0 \le v \le 1$$

Hence deduce that the confidence limits for θ corresponding to confidence coefficient α are L and $\dfrac{L}{(1-\alpha)^{1/n}}$

[Delhi Univ. B.Sc. (Stat. Hons.), 1982, 1983]

Solution. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the population (*) and let $L = \max(X_1, X_2, \ldots, X_n)$. The distribution of L is given by $\quad dG(L) = n[F(L)]^{n-1} . f(L)\, dL$

where $F(.)$ is the distribution function of X given by

$$F(L) = \int_0^L f(x, \theta) dx = \frac{L}{\theta}$$

$$\therefore \qquad dG(L) = n\left(\frac{L}{\theta}\right)^{n-1} \cdot \frac{dL}{\theta}, \; 0 \le L \le \theta$$

If we take $V = L/\theta$, the Jacobian of transformation is θ. Hence p.d.f. $h(.)$ of V is given by

$$h(v) = nv^{n-1} . \frac{1}{\theta} \mid J \mid = nv^{n-1}, \; 0 \le v \le 1$$

which is independent of θ.

To obtain the confidence limits for θ, with confidence coefficient α, let us define v_α such that

$$P(v_\alpha < V < 1) = \alpha \quad \Rightarrow \quad \int_{v_\alpha}^1 h(v)\, dv = \alpha \qquad \ldots(**)$$

$$\Rightarrow \quad n \int_{v_\alpha}^1 v^{n-1}\, dv = \alpha \quad \Rightarrow \quad 1 - v_\alpha^{\,n} = \alpha$$

$$\Rightarrow \quad v_\alpha = (1-\alpha)^{1/n} \qquad \ldots(***)$$

From (**) and (***), we get

$$P[(1-\alpha)^{1/n} < V < 1] = \alpha$$

$$\Rightarrow \quad P\left[(1-\alpha)^{1/n} < \frac{L}{\theta} < 1\right] = \alpha$$

$$\Rightarrow \quad P\left[L < \theta < \frac{L}{(1-\alpha)^{1/n}}\right] = \alpha$$

Hence the required confidence limits for θ are L and $L/(1-\alpha)^{1/n}$.

Example 15·47. *Given a random sample from a population with p.d.f.*

$$f(x, \theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta$$

show that 100 (1 – α)% confidence interval for θ is given by R and R/ψ where ψ is given by

$$\psi^{n-1}[n-(n-1)\psi] = \alpha,$$

and R is the sample range.

Solution. The joint p.d.f. of $x_1, x_2, \ldots, x_n$ is given by

$$L = \frac{1}{\theta^n}, \quad 0 \le x_i \le \theta$$

If $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is the ordered sample then the joint p.d.f. of $x_{(n)}$ and $x_{(1)}$ is given by

$$g(x_{(1)}, x_{(n)}) = \frac{n(n-1)}{\theta^n}\left[x_{(n)} - x_{(1)}\right]^{n-2}, \quad 0 \le x_{(1)} \le x_{(n)} \le \theta$$

To obtain the distribution of the sample range R, let us make the transformation of variables

$$R = x_{(n)} - x_{(1)} \text{ and } v = x_{(1)} \quad \Rightarrow \quad v = x_{(n)} - R \le \theta - R$$

The Jacobian of transformation is $|J| = 1$ and the joint p.d.f. of R and V becomes

$$h(R, v) = \frac{n(n-1)}{\theta^n} R^{n-2}, \quad 0 < v < \theta - R$$

The marginal density of R is given by

$$h_1(R) = \int_0^{\theta-R} \frac{n(n-1)}{\theta^n} \cdot R^{n-2}\, dv$$

$$= \frac{n(n-1)\, R^{n-2}(\theta - R)}{\theta^n}, \quad 0 \le R \le \theta$$

The density of $U = R/\theta$ is

$$h_2(u) = h_1(R)\left|\frac{dR}{du}\right| = \frac{n(n-1)\, R^{n-2}(\theta - R)}{\theta^n} \cdot \theta$$

$$= n(n-1)u^{n-2}(1-u), \quad 0 \le u \le 1$$

100 (1 – α)% confidence interval for θ is given by

$$P(\psi \le U \le 1) = 1 - \alpha \qquad \ldots(*)$$

where ψ is obtained from the equation

$$\int_0^{\psi} h_2(u) d = \alpha$$

$$\Rightarrow \quad n(n-1)\int_0^{\psi} u^{n-2}(1-u)du = \alpha$$

$$\Rightarrow \quad \left| nu^{n-1} - (n-1)\, u^n \right|_0^{\psi} = \alpha$$

$$\Rightarrow \quad \psi^{n-1}[n-(n-1)\psi] = \alpha \qquad \ldots(**)$$

From (*), we get

$$P\left[\psi \le \frac{R}{\theta} \le 1\right] = 1 - \alpha$$

$$\Rightarrow \qquad P\left[R \le \theta \le \frac{R}{\psi}\right] = 1 - \alpha$$

Hence the required limits for θ are given by R and R/ψ, where ψ is the solution of (**).

Example 15·48. *Given one observation from a population with p.d.f.*

$$f(x, \theta) = \frac{2}{\theta^2}(\theta - x), \quad 0 \le x \le \theta,$$

obtain 100 (1 – α)% confidence interval for θ.

[Delhi Univ. B.Sc. (Stat Hons.), 1991]

Solution. The density of $u = x/\theta$ is given by

$$g(u) = f(x, \theta) \cdot \left| \frac{dx}{du} \right| = \frac{2}{\theta^2}(\theta - x).\, \theta$$

$$= 2(1 - u), \quad 0 \le u \le 1$$

To obtain 100 (1 – α)% confidence interval for θ, we choose two quantities u_1 and u_2 such that

$$P[u_1 \le u \le u_2] = 1 - \alpha \qquad \ldots(*)$$

and $$P[u < u_1] = P[u > u_2] = \alpha/2$$

Now $$P[u < u_1] = \frac{\alpha}{2} \Rightarrow \int_0^{u_1} 2(1 - u)\, du = \frac{\alpha}{2}$$

$$\Rightarrow \qquad u_1^2 - 2u_1 + \frac{\alpha}{2} = 0 \qquad \ldots(**)$$

Similarly, $$P(u > u_2) = \frac{\alpha}{2} \Rightarrow \int_{u_2}^{1} 2(1 - u)\, du = \frac{\alpha}{2}$$

$$\Rightarrow \qquad u_2^2 - 2u_2 + \left(1 - \frac{\alpha}{2}\right) = 0 \qquad \ldots(***)$$

From (*), we get

$$P\left[u_1 \le \frac{x}{\theta} \le u_2\right] = 1 - \alpha \Rightarrow P\left[\frac{x}{u_2} \le \theta \le \frac{x}{u_1}\right] = 1 - \alpha$$

Hence the required interval for θ is $\left(\frac{x}{u_2}, \frac{x}{u_1}\right)$, where u_1 and u_2 are given by (**) and (***).

15·15·1. Confidence Intervals for Large Samples. It has been proved that under certain regularity conditions, the first derivative of the logarithm of the likelihood function w.r.t parameter θ *viz.*, $\frac{\partial}{\partial \theta} \log L$, is asymptotically normal with mean zero and variance given by

$$\text{Var}\left(\frac{\partial}{\partial\theta}\log L\right)=E\left(\frac{\partial}{\partial\theta}\log L\right)^2=E\left(-\frac{\partial^2}{\partial\theta^2}\log L\right)$$

Hence for large n,

$$Z=\frac{\dfrac{\partial}{\partial\theta}\log L}{\sqrt{\text{Var}\left(\dfrac{\partial}{\partial\theta}\log L\right)}}\sim N(0,1) \qquad \text{...(15·68)}$$

The result enables us to obtain confidence interval for the parameter θ in large samples. Thus for large samples, the confidence interval for θ with confidence coefficient $(1-\alpha)$ is obtained by converting the inequalities in

$$P[|Z|\le\lambda_\alpha]=1-\alpha \qquad \text{...(15·69)}$$

where λ_α is given by

$$\frac{1}{\sqrt{2\pi}}\int_{-\lambda_\alpha}^{\lambda_\alpha}\exp(-u^2/2)\,du=1-\alpha \qquad \text{...[15·69(a)]}$$

Example 15·49. *Obtain 100 $(1-\alpha)$% confidence limits (for large samples) for the parameter λ of the Poisson distribution*

$$f(x,\lambda)=\frac{e^{-\lambda}\,.\,\lambda^x}{x\,!}\;;\; x=0,1,2,\ldots$$

Solution. We have

$$\frac{\partial}{\partial\lambda}\log L=\frac{\partial}{\partial\lambda}\left[-n\lambda+\left(\sum_{i=1}^{n}x_i\right)\log\lambda-\sum_{i=1}^{n}\log x_i\right]$$

$$=-n+\frac{\Sigma x_i}{\lambda}=n\left(\frac{\bar{x}}{\lambda}-1\right)$$

$$\text{Var}\left(\frac{\partial}{\partial\lambda}\log L\right)=E\left(-\frac{\partial^2}{\partial\lambda^2}\log L\right)=E\left(\frac{n\bar{x}}{\lambda^2}\right)$$

$$=\frac{n}{\lambda^2}E(\bar{x})=\frac{n}{\lambda}$$

$$\therefore\quad Z=\frac{n\left(\dfrac{\bar{x}}{\lambda}-1\right)}{\sqrt{n/\lambda}}=\sqrt{(n/\lambda)}\,(\bar{x}-\lambda)\sim N(0,1) \qquad \text{[Using (15·68)]}$$

Hence 100 $(1-\alpha)$% confidence interval for λ is given by (for large samples)

$$P\left[|\sqrt{(n/\lambda)}\,(\bar{x}-\lambda)|\le\lambda_\alpha\right]=1-\alpha$$

Hence the required limits for λ are the roots of the equation :

$$|\sqrt{n/\lambda}\,(\bar{x}-\lambda)|=\lambda_\alpha$$

$$\Rightarrow\qquad n(\bar{x}-\lambda)^2-\lambda\,.\,\lambda_\alpha^2=0$$

$$\Rightarrow \qquad \lambda^2 - \lambda\left(2\,\bar{x} + \frac{\lambda_\alpha^2}{n}\right) + \bar{x}^2 = 0$$

$$\Rightarrow \quad \lambda = \frac{\left(2\,\bar{x} + \frac{\lambda_\alpha^2}{n}\right) \pm \left[\left(2\,\bar{x} + \frac{\lambda_\alpha^2}{n}\right)^2 - 4\,\bar{x}^2\right]^{1/2}}{2} \qquad \ldots(*)$$

For example, 95% confidence interval for λ is given by taking $\lambda_\alpha = 1{\cdot}96$ in (*), thus giving

$$\lambda = \frac{1}{2}\left(2\,\bar{x} + \frac{3{\cdot}84}{n}\right) \pm \sqrt{\left(\frac{3{\cdot}84\bar{x}}{n} + \frac{3{\cdot}69}{n^2}\right)} = \bar{x} \pm 1{\cdot}96\sqrt{\frac{\bar{x}}{n}},$$

to order $n^{-1/2}$.

Example 15·50. *Show that for the distribution :*

$$dF(x) = \theta\, e^{-x\theta}\,;\, 0 < x < \infty$$

central confidence limits for θ for large samples with 95% confidence coefficient are given by

$$\theta = \left(1 \pm \frac{1{\cdot}96}{\sqrt{n}}\right) \Big/ \bar{x}$$

Solution. Here
$$L = \theta^n \exp\left[-\theta \sum_{i=1}^{n} x_i\right]$$

$$\frac{\partial}{\partial\theta}\log L = \frac{\partial}{\partial\theta}\left[n \log\theta - \theta\,\Sigma\, x_i\right]$$

$$= \frac{n}{\theta} - \sum_{i=1}^{n} x_i = n\left(\frac{1}{\theta} - \bar{x}\right)$$

$$\frac{\partial^2}{\partial\theta^2}\log L = -\frac{n}{\theta^2}$$

$$\therefore \qquad \text{Var}\left(\frac{\partial}{\partial\theta}\log L\right) = E\left(-\frac{\partial^2}{\partial\theta^2}\log L\right) = \frac{n}{\theta^2}$$

Hence, for large samples, using (15·68) we have :

$$Z = \frac{n\left(\frac{1}{\theta} - \bar{x}\right)}{\sqrt{n/\theta^2}} \sim N(0, 1) \;\Rightarrow\; \sqrt{n}\,(1 - \theta\,\bar{x}) \sim N(0, 1)$$

Hence 95% confidence limits for θ are given by

$$P[-1{\cdot}96 \le \sqrt{n}\,(1 - \theta\,\bar{x}) \le 1{\cdot}96] = 0{\cdot}95$$

Now $\quad \sqrt{n}\,(1 - \theta\bar{x}) \le 1{\cdot}96 \;\Rightarrow\; \left(1 - \dfrac{1{\cdot}96}{\sqrt{n}}\right)\dfrac{1}{\bar{x}} \le \theta \qquad \ldots(*)$

and $\quad -1{\cdot}96 \le \sqrt{n}\,(1 - \theta\bar{x}) \;\Rightarrow\; \theta \le \left(1 + \dfrac{1{\cdot}96}{\sqrt{n}}\right)\dfrac{1}{\bar{x}} \qquad \ldots(**)$

Hence, from (*) and (**), the central 95% confidence limits for θ are given by

$$\theta = \left(1 \pm \frac{1{\cdot}96}{\sqrt{n}}\right) \cdot \frac{1}{\bar{x}}$$

EXERCISE 15 (c)

1. Discuss the concept of interval estimation and provide suitable illustration. *[Delhi Univ. M.A. (Eco.), 1987]*

2. Critically examine how interval estimation differs from point estimation. Give the 95% confidence interval for the mean of the normal distribution, when its variance is known.

[Madras Univ. B.Sc. Sept., 1988]

3. What are confidence intervals ? How are they constructed using t-distribution ?

[Madras Univ. B.Sc., March, 1989]

4. The random variable X is uniformly distributed in $(a, a + 2)$. Obtain limits x_1 and x_2 such that

$$P(X \le x_1) = P(X \ge x_2) = 0{\cdot}025$$

The random variable is observed once, the value being x_0. Give a method of obtaining an interval estimate for 'a' which you expect to be correct in 95% of trials. *[Calcutta Univ. B.Sc. (Maths. Hons.), 1990]*

5. Obtain 100 $(1 - \alpha)$% confidence interval *either* for the unknown parameter p of a binnomial distribution when the parameter n is known *or* for the population correlation coefficient when the population is Normal.

[Delhi Univ. B.Sc. (Stat. Hons.), 1983]

6. Let $f_\theta(x) = 1/\theta,\ 0 \le x \le \theta$ and let L be the largest observation of a sample of size n from the above distribution.

Obtain the distribution of (L/θ) and hence deduce that the confidence limits corresponding to confidence coefficient α are L, and $\dfrac{L}{(1-\alpha)^{1/n}}$ respectively.

[Delhi Univ. B.Sc. (Stat. Hons.), 1992

7. (*a*) What are confidence intervals ? y is the largest observation in a sample of size n drawn from a rectangular population in $(0, \theta)$. Find the confidence coefficient corresponding to the confidence interval

$$\left\{y,\ y/(1-\alpha)^{1/n}\right\}$$

where 'α' is the level significance.

[Bhartiyan Univ. M.Sc. (Maths.), 1991]

(*b*) Prove that the confidence interval for θ obtained in (*a*) part above is shorter than the one obtained in Question 9 below.

8. Develop a general method for constructing confidence intervals. Consider a random sample of size n from the exponential distribution with p.d.f.

$$f(x, \theta) = e^{-(x-\theta)},\ \theta \le x < \infty,\ -\infty < \theta < \infty.$$

Show that $P\left[X_{(1)} - \frac{1}{n}\log\alpha \le \theta \le X_{(1)}\right] = 1 - \alpha$

where symbols have their usual meanings. Also interpret the result.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

9. Consider a random sample $X_1, X_2, \ldots, X_n$ from an $U[0, \theta]$ population. Show that R and R/ξ are the confidence limits for θ with confidence coefficient $(1-\alpha)$, where R is the sample range and ξ satisfies the equation :

$$\xi^{n-1}\{n-(n-1)\xi\} = \alpha$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1985]

10. Explain the difference between point estimation and interval estimation.

Obtain 100 $(1-\alpha)$% confidence interval for the population correlation coefficient 'ρ' when the random sample of size n has been drawn from bivariate normal population.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

11. Describe the pivotal quantity method for constructing confidence intervals.

Obtain a large sample 100 $(1-\alpha)$% confidence interval for the parameter θ in random sampling from the population :

$$dF(x) = \theta\, e^{-\theta x}\,;\, x > 0,\ \theta > 0$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

12. Develop a general method for obtaining confidence intervals. Obtain a $100(1-\alpha)$% confidence interval for large sample size for the parameter θ of the Poisson distribution :

$$f(x, \theta) = \frac{e^{-\theta}\,\theta^x}{x\,!}, x = 0, 1, 2, \ldots$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

13. Describe the general method of constructing the confidence interval for large samples.

If $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with mean θ, obtain 95% confidence interval for θ when n is large.

[Delhi Univ. B.Sc. (Stat. Hons.), 1993]

14. (*a*) Show that with the exponential distribution

$$dF(x) = \theta e^{-\theta x},\ x \ge 0$$

central confidence limits for θ for large samples of size n and 95% confidence coefficient are :

$$\{1 \pm 1{\cdot}96/\sqrt{n}\}/\bar{x}\,,$$

where $\bar{x}$ is the mean of the sample observations $x_1, x_2, \ldots, x_n$ drawn randomly from the exponential population.

[Indian Civil Services, 1983]

(*b*) Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with density function : $f(x, \theta) = \theta e^{-\theta x}, 0 \le x < \infty$

Find a $100(1-\alpha)$ (when $0 < \alpha < 1$) percent confidence interval for the mean of this population, for large samples.

[Madras Univ. B.Sc., 1991]

15. (*a*) Discuss the problem of interval estimation. Obtain the minimum confidence interval for the variance for a random sample of size n from a normal population with unknown mean.

[Indian Civil Services, 1991]

(*b*) Give a method of determining the confidence limits for a single unknown parameter, stating the conditions of validity. From amongst intervals of Confidence Coefficient α, how will you decide one as being superior to another ?

[Indian Civil Services, 1989]

16. Consider a random sample $X_1, X_2, \ldots, X_n$ from the exponential distribution with p.d.f.

$$
\begin{aligned}
f(x, \theta, p) &= \frac{\exp(-x/\theta) \cdot x^{p-1}}{\Gamma p \; \theta^p}, \quad x > 0 \\
&= 0 \qquad\qquad\qquad\quad , \text{ otherwise}
\end{aligned}
$$

If p is known, obtain a confidence interval for θ, starting from the sufficient statistic $\overline{X}/p$.

CHAPTER SIXTEEN

Statistical Inference-II
(Testing of Hypothesis, Non-parametric Methods and Sequential Analysis)

16·1. Introduction. The main problems in statistical inference can be broadly classified into two areas :

(*i*) The area of estimation of population parameters and setting up of confidence intervals for them, *i.e.*, the area of *point and interval estimation* and

(*ii*) *Tests of statistical hypothesis.*

The first topic has already been discussed in Chapter 15. In this chapter we shall discuss: (*a*) The theory of testing of hypothesis initiated by J. Neyman and E.S. Pearson (Section 16·2), (*b*) Sequential analysis propounded by A. Wald (Section 16·4) and (*c*) Non-parametric tests (Section 16·3). In Neyman-Pearson theory, we use statistical methods to arrive at decisions in certain situations where there is lack of certainty, on the basis of a sample whose size is fixed in advance while in Wald's sequential theory the sample size is not fixed but is regarded as a random variable. Before taking up a detailed discussion of the topics in (*a*), (*b*) and (*c*), we shall explain below certain concepts which are of fundamental importance.

16·2. Statistical Hypothesis-Simple and Composite. *A statistical hypothesis is some statement or assertion about a population or equivalently about the probability distribution characterising a population which we want to verify on the basis of information available from a sample.* If the statistical hypothesis specifies the population completely then it is termed as a *simple statistical hypothesis*, otherwise it is called a *composite statistical hypothesis.*

For example, if $X_1, X_2, ..., X_n$ is a random sample of size n from a normal population with mean μ and variance σ^2, then the hypothesis

$$H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$$

is a simple hypothesis, whereas each of the following hypotheses is a composite hypothesis:

(*i*) $\mu = \mu_0$, (*ii*) $\sigma^2 = \sigma_0^2$

(*iii*) $\mu < \mu_0, \sigma^2 = \sigma_0^2$ (*iv*) $\mu > \mu_0, \sigma^2 = \sigma_0^2$

(*v*) $\mu = \mu_0, \sigma^2 < \sigma_0^2$, (*vi*) $\mu = \mu_0, \sigma^2 > \sigma_0^2$

(*vii*) $\mu < \mu_0, \sigma^2 > \sigma_0^2$.

A hypothesis which does not specify completely 'r' parameters of a population is termed as a *composite hypothesis with r degrees of freedom.*

16·2·1. Test of a Statistical Hypothesis. *A test of a statistical hypothesis is a two-action decision problem after the experimental sample values have been obtained, the two-actions being the acceptance or rejection of the hypothesis under consideration.*

16·2·2. Null Hypothesis. In hypothesis testing, a statistician or decision-maker should not be motivated by prospects of profit or loss resulting from the acceptance or rejection of the hypothesis. He should be completely impartial and should have no brief for any party or company nor should he allow his personal views to influence the decision. *Much, therefore, depends upon how the hypothesis is framed.* For example, let us consider the 'light-bulbs' problem. Let us suppose that the bulbs manufactured under some standard manufacturing process have an average life of μ hours and it is proposed to test a new procedure for manufacturing light bulbs. Thus, we have two populations of bulbs, those manufactured by standard process and those manufactured by the new process. In this problem the following three hypotheses may be set up :

(*i*) New process is better than standard process.

(*ii*) New process is inferior to standard process.

(*iii*) There is no difference between the two processes.

The first two statements appear to be biased since they reflect a preferential attitude to one or the other of the two processes. Hence the best course is to adopt *the hypothesis of no difference,* as stated in (*iii*). This suggests that the *statistician should take up the neutral or null attitude regarding the outcome of the test.* His attitude should be on the null or zero line in which the experimental data has the due importance and complete say in the matter. *This neutral or non-committal attitude of the statistician or decision-maker before the sample observations are taken is the keynote of the null hypothesis.*

Thus, in the above example of light bulbs if μ_0 is the mean life (in hours) of the bulbs manufactured by the new process then the null hypothesis which is usually denoted by H_0, can be stated as follows :

$$H_0 : \mu = \mu_0.$$

As another example let us suppose that two different concerns manufacture drugs for inducing sleep, drug A manufactured by first concern and drug B manufactured by second concern. Each company claims that its drug is superior to that of the other and it is desired to test which is a superior drug A or B ? To formulate the statistical hypothesis let X be a random variable which denotes the additional hours of sleep gained by an individual when drug A is given and let the random variable Y denote the additional hours of sleep gained when drug B is used. Let us suppose that X and Y follow the probability distributions with means μ_X and μ_Y respectively. Here our *null hypothesis would be that there is no difference between the effects of two drugs.* Symbolically,

$$H_0 : \mu_X = \mu_Y.$$

16·2·3. Alternative Hypothesis. It is desirable to state what is called an *alternative hypothesis* in respect of every statistical hypothesis being tested because the acceptance or rejection of null hypothesis is meaningful only when it is being tested against a rival hypothesis which should rather be explicitly mentioned. *Alternative hypothesis* is usually denoted by H_1. For example, in the example of light bulbs, alternative hypothesis could be $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$. In the example of drugs, the alternative hypothesis could be $H_1 : \mu_X > \mu_Y$ or $\mu_X < \mu_Y$ or $\mu_X \neq \mu_Y$.

In both the cases, the first two of the alternative hypotheses give rise to what are called '*one tailed*' tests and the third alternative hypothesis results in '*two tailed*' tests.

Important Remarks 1. In the formulation of a testing problem and devising a 'test of hypothesis' the roles of H_0 and H_1 are not at all symmetric. In order to decide which one of the two hypotheses should be taken as null hypothesis H_0 and which one as alternative hypothesis H_1, the intrinsic difference between the roles and the implifications of these two terms should be clearly understood.

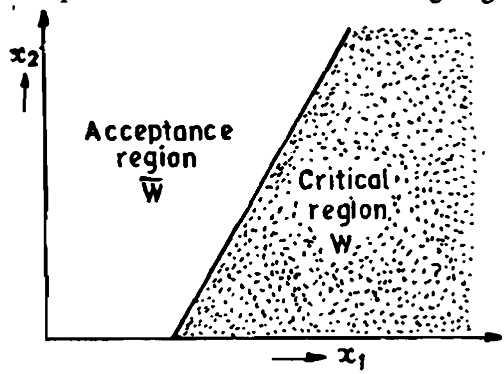
2. If a particular problem cannot be stated as a test between two simple hypotheses, *i.e.*, simple null hypothesis against a simple alternative hypothesis, then the next best alternative is to formulate the problem as the test of a simple null hypothesis against a composite alternative hypothesis. In other words, one should try to structure the problem so that null hypothesis is simple rather than composite.

3. Keeping in mind the potential losses due to wrong decisions (which may or may not be measured in terms of money), the decision maker is somewhat conservative in holding the null hypothesis as true unless there is a strong evidence from the experimental sample observations that it is false. To him, the *consequences of wrongly rejecting a null hypothesis seem to be more severe than those of wrongly accepting it.* In mot of the cases, the statistical hypothesis is in the form of a claim that a particular product or product process is superior to some existing standard. The null hypothesis H_0 in this case is that there is no difference between the new product or production process and the existing standard. In other words, null hypothesis nullifies this claim. The rejection of the null hypothesis wrongly which amounts to the acceptance of claim wrongly involves huge amount of pocket expenses towards a substantive overhaul of the existing set-up. The resulting loss is comparatively regarded as more serious than the opportunity loss in wrongly accepting H_0 which amounts to wrongly rejecting the claim, *i.e.*, in sticking to the less efficient existing standard. In the light-bulbs problem discussed earlier, suppose the research division of the concern, on the basis of the limited experimentation, claims that its brand is more effective than that manufactured by standard process. If in fact, the brand fails to be more effective the loss incurred by the concern due to an immediate obsolescence of the product, decline of the concern's image, etc., will be quite serious. On the other hand, the failure to bring out a superior brand in the market is an opportunity loss and is not a consideration to be as serious as the other loss.

16·2·4. Critical Region. Let $x_1, x_2, \ldots, x_n$ be the sample observations denoted by O. All the values of O will be aggregate of a sample and they constitute a space, called the *sample space*, which is denoted by S.

Since the sample values $x_1, x_2, \ldots, x_n$ can be taken as a point in n-dimensional space, we specify some region of the n-dimensional space and see whether this point lies within this region or outside this region. We divide the whole sample space S into two disjoint parts W and $S - W$ or $\bar{W}$ or W'. The null hypothesis H_0 is rejected if the observed sample point falls in W and if it falls in W' we reject H_1 and accept H_0. *The region of rejection of H_0 when H_0 is true is that region of the outcome set where H_0 is rejected if the sample point falls in that region and is called critical region.* Evidently, the size of the critical region is α, the probability of committing type 1 error (discussed below).

Suppose if the test is based on a sample of size 2, then the outcome set or the sample space is the first quadrant in a two-dimensional space and a test criterion will enable us to separate our outcome set into two complementary subsets, W and $\bar{W}$. If the sample point falls in the subset W, H_0 is rejected, otherwise H_0 is accepted. This is shown in the following diagram :



16·2·5. Two Types of Errors. The decision to accept or reject the null hypothesis H_0 is made on the basis of the information supplied by the observed sample observations. The conclusion drawn on the basis of a particular sample may not always be true in respect of the population. The four possible situations that arise in any test procedure are given in the following table.

DOUBLE DICHOTOMY RELATING TO DECISION AND HYPOTHESIS

		Decision From Sample	
		Reject H_0	*Accept H_0*
True State	*H_0 True*	Wrong (Type I Error)	Correct
	H_0 False (*H_1 True*)	Correct	Wrong (Type II Error)

From the above table it is obvious that in any testing problem we are liable to commit two types of errors.

Errors of Type I and Type II. The error of rejecting H_0 (accepting H_1) when H_0 is true is called *Type I error* and the error of accepting H_0 when H_0 is false (H_1 is true) is called *Type II* error. The probabilities of type I and type II errors are denoted by α and β respectively. Thus

$$\begin{aligned} \alpha &= \text{Probability of type I error} \\ &= \text{Probability of rejecting } H_0 \text{ when } H_0 \text{ is true.} \\ \beta &= \text{Probability of type II error} \\ &= \text{Probability of accepting } H_0 \text{ when } H_0 \text{ is false.} \end{aligned}$$

Symbolically:

$$\left.\begin{aligned} & P(\mathbf{x} \in W \mid H_0) = \alpha, \quad \text{where } \mathbf{x} = (x_1, x_2, \ldots, x_n) \\ \Rightarrow \quad & \int_W L_0 \, d\mathbf{x} = \alpha \end{aligned}\right\} \qquad \ldots(16{\cdot}1)$$

where L_0 is the likelihood function of the sample observations under H_0 and $\int d\mathbf{x}$ represents the n-fold integral

$$\int\int \ldots \int dx_1 \, dx_2 \ldots dx_n.$$

Again

$$\left.\begin{aligned} & P(\mathbf{x} \in \overline{W} \mid H_1) = \beta \\ \Rightarrow \quad & \int_{\overline{W}} L_1 \, d\mathbf{x} = \beta \end{aligned}\right\} \qquad \ldots(16{\cdot}2)$$

where L_1 is the likelihood function of the sample observations under H_1. Since

$$\int_W L_1 \, d\mathbf{x} + \int_{\overline{W}} L_1 \, d\mathbf{x} = 1,$$

we get

$$\int_W L_1 \, d\mathbf{x} = 1 - \int_{\overline{W}} L_1 \, d\mathbf{x} = 1 - \beta \qquad \ldots(16{\cdot}2a)$$

$$\Rightarrow \quad P(\mathbf{x} \in W \mid H_1) = 1 - \beta \qquad \ldots(16{\cdot}2b)$$

16·2·6. Level of Significance. α, the probability of type 1 error, is known as the level of significance of the test. It is also called the *size of the critical region*.

16·2·7. Power of the Test. $1 - \beta$, defined in (16·2*a*) and (16·2*b*) is called the *power function of the test hypothesis H_0 against the alternative hypothesis H_1*. The value of the power function at a parameter point is called the *power of the test* at that point.

Remarks 1. In quality control terminology, α and β are termed as *producer's risk and consumer's risk*, respectively.

2. An ideal test would be the one which properly keeps under control both the types of errors. But since the commission of an error of either type is a random variable, equivalently an ideal test should minimise the probability of both the types of errors, *viz.*, α and β. But unfortunately, for a fixed sample size n, α and β are so related (like producer's and consumer's risk in sampling inspection plans), that the reduction in one results in an increase in the other. Consequently, the simultaneous minimising of both the errors is not possible.

Since type I error is deemed to be more serious than the type II error (c.f. Remark 3, § 16·2·3) the usual practice is to control α at a predetermined low level and subject to this constraint on the probabilities of type I error, *choose a test which minimises* β *or maximises the power function* $1-\beta$. Generally, we choose $\alpha = 0{\cdot}05$ or $0{\cdot}01$.

16·3. Steps in Solving Testing of Hypothesis Problem. The major steps involved in the solution of a 'testing of hypothesis' problem may be outlined as follows :

1. Explicit knowledge of the nature of the population distribution and the parameter(s) of interest, *i.e.*, the parameter(s) about which the hypotheses are set up.

2. Setting up of the null hypothesis H_0 and the alternative hypothesis H_1 in terms of the range of the parameter values each one embodies.

3. The choice of a suitable statistic $t = t(x_1, x_2, \ldots, x_n)$ called the *test statistic*, which will best reflect upon the probability of H_0 and H_1.

4. Partitioning the set of possible values of the test statistic t into two disjoint sets W (called the *rejection region* or *critical region*) and $\overline{W}$ (called the *acceptance region*) and framing the following test :

(*i*) Reject H_0 (*i.e.*, accept H_1) if the value of t falls in W.

(*ii*) Accept H_0 if the value of t falls in $\overline{W}$.

5. After framing the above test, obtain experimental sample observations, compute the appropriate test statistic and take action accordingly.

16·4. Optimum Test Under Different Situations. The discussion in § 16·3 and Remark 2, § 16·2·6 enables us to obtain the so called best test under different situations. In any testing problem the first two steps, *viz.*, the form of the population distribution, the parameter(s) of interest and the framing of H_0 and H_1 should be obvious from the description of the problem. The most crucial step is the choice of the '*best test*, *i.e.*, the best statistic 't' and the critical region W where *by best test we mean one which in addition to controlling* α *at any desired low level has the minimum type II error* β *or maximum power* $1-\beta$, *compared to* β *of all other tests having this* α. This leads to the following definition.

16·4·1. Most Powerful Test (MP Test). Let us consider the problem of testing a simple hypothesis

$$H_0 : \theta = \theta_0$$

against a simple alternative hypothesis

$$H_1 : \theta = \theta_1$$

Definition. *The critical region W is the most powerful (MP) critical region of size* α *(and the corresponding test a most powerful test of level* α*) for testing* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta = \theta_1$ *if*

$$P(\mathbf{x} \in W \mid H_0) = \int_W L_0 \, d\mathbf{x} = \alpha \qquad \ldots (16{\cdot}3)$$

and
$$P(\mathbf{x} \in W \mid H_1) \geq P(\mathbf{x} \in W_1 \mid H_1) \qquad \ldots (16{\cdot}3a)$$

for every other critical region W_1 satisfying (16·3).

16·4·2. Uniformly Most Powerful Test (UMP Test). Let us now take up the case of testing a simple null hypothesis against a composite alternative hypothesis, *e.g.*, of testing

$$H_0 : \theta = \theta_0$$

against the alternative

$$H_1 : \theta \neq \theta_0$$

In such a case, for a predetermined α, the best test for H_0 is called the *uniformly most powerful test of level* α.

Definition. *The region W is called uniformly most powerful (UMP) critical region of size α [and the corresponding test as uniformly most powerful (UMP) test of level α] for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ i.e., $H_1 : \theta = \theta_1 \neq \theta_0$ if*

$$P(\mathbf{x} \in W \mid H_0) = \int_W L_0 \, dx = \alpha \qquad \ldots(16\cdot4)$$

and $$P(\mathbf{x} \in W \mid H_1) \geq P(\mathbf{x} \in W_1 \mid H_1) \text{ for all } \theta \neq \theta_0, \qquad \ldots(16\cdot4a)$$

whatever the region W_1 satisfying (16·4) may be.

16·5. Neyman J. and Pearson, E.S. Lemma. This Lemma provides the most powerful test of simple hypothesis against a simple alternative hypothesis. The theorem, known as Neyman-Pearson Lemma, will be proved for density function $f(x, \theta)$ of a single continuous variate and a single parameter. However, by regarding x and θ as vectors, the proof can be easily generalised for any number of random variables $x_1, x_2, \ldots, x_n$ and any number of parameters $\theta_1, \theta_2, \ldots, \theta_k$. The variables $x_1, x_2, \ldots, x_n$ occurring in this theorem are understood to represent a random sample of size n from the population whose density function is $f(x, \theta)$. The lemma is concerned with a simple hypothesis $H_0 : \theta = \theta_0$ and a simple alternative $H_1 : \theta = \theta_1$.

Theorem 16·1. (*Neyman-Pearson Lemma*). *Let $k > 0$, be a constant and W be a critical region of size α such that*

$$W = \left\{ \mathbf{x} \in S : \frac{f(x, \theta_1)}{f(x, \theta_0)} > k \right\}$$

$$\Rightarrow \qquad W = \left\{ \mathbf{x} \in S : \frac{L_1}{L_0} > k \right\} \qquad \ldots(16\cdot5)$$

and $$\overline{W} = \left\{ \mathbf{x} \in S : \frac{L_1}{L_0} \leq k \right\} \qquad \ldots(16\cdot5a)$$

where L_0 and L_1 are the likelihood functions of the sample observations $x = (x_1, x_2, \ldots, x_n)$ under H_0 and H_1 respectively. Then W is the most powerful critical region of the test hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$.

Proof. We are given

$$P(\mathbf{x} \in W \mid H_0) = \int_W L_0 \, d\mathbf{x} = \alpha \qquad \ldots(16\cdot6)$$

The power of the region is

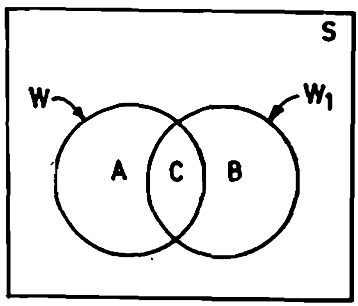
$$P(\mathbf{x} \in W \mid H_1) = \int_W L_1 \, d\mathbf{x} = 1 - \beta, \text{ (say).} \qquad \text{... (16·6a)}$$

In order to establish the lemma, we have to prove that there exists no other critical region, of size less than or equal to α, which is more powerful than W. Let W_1 be another critical region of size $\alpha_1 \leq \alpha$ and power $1 - \beta_1$ so that we have

$$P(\mathbf{x} \in W_1 \mid H_0) = \int_{W_1} L_0 \, d\mathbf{x} = \alpha_1 \qquad \text{.... (16·7)}$$

and $$P(\mathbf{x} \in W_1 \mid H_1) = \int_{W_1} L_1 \, d\mathbf{x} = 1 - \beta_1 \qquad \text{... (16·7a)}$$

Now we have to prove that $1 - \beta \geq 1 - \beta_1$



Let $W = A \cup C$ and $W_1 = B \cup C$

(C may be empty, *i.e.*, W and W_1 may be disjoint).

If $\alpha_1 \leq \alpha$, we have

$$\int_{W_1} L_0 \, d\mathbf{x} \leq \int_W L_0 \, d\mathbf{x}$$

$$\Rightarrow \quad \int_{B \cup C} L_0 \, d\mathbf{x} \leq \int_{A \cup C} L_0 \, d\mathbf{x}$$

$$\Rightarrow \quad \int_B L_0 \, d\mathbf{x} \leq \int_A L_0 \, d\mathbf{x}$$

$$\Rightarrow \quad \int_A L_0 \, d\mathbf{x} \geq \int_B L_0 \, d\mathbf{x} \qquad \text{... (16·8)}$$

Since $A \subset W$,

$$\text{(16·5)} \Rightarrow \int_A L_1 \, d\mathbf{x} > K \int_A L_0 \, d\mathbf{x} \geq k \int_B L_0 \, d\mathbf{x} \qquad \text{... (16·8a)}$$

[Using (16·8)]

Also [16·5 (*a*)] implies

$$\frac{L_1}{L_0} \le k \ \forall x \in \overline{W}$$

$$\Rightarrow \qquad \int_{\overline{W}} L_1 \ dx \ \le k \int_{\overline{W}} L_0 \ dx$$

This result also holds for any subset of $\overline{W}$, say $\overline{W} \cap W_1 = B$. Hence

$$\int_B L_1 \ dx \ \le k \int_B L_0 \ dx \le \int_A L_1 \ dx \quad \text{[From (16·8}a\text{)]}$$

Adding $\int_C L_1 \ dx$ to both sides, we get

$$\int_{W_1} L_1 \ dx \ \le \int_W L_1 \, dx$$

$$\Rightarrow \qquad 1-\beta \ \ge 1-\beta_1$$

Hence the Lemma.

Remark. Let W defined in (16·5) of the above theorem be the most powerful critical region of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, and let it be independent of $\theta_1 \in \Theta_1 = \Theta - \Theta_0$, where Θ_0 is the parameter space under H_0. Then we say that C.R. W is the UMP CR of size α for testing $H_0 : \theta = \theta_0$, against $H_1 : \theta \in \Theta_1$.

16·5·1. Unbiased Test and Unbiased Critical Region. Let us consider the testing of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. *The critical region W and consequently the test based on it is said to be unbiased if the power of the test exceeds the size of the critical region, i.e., if*

$$\text{Power of the test } \ge \text{ size of the C.R.} \qquad \dots (16.9)$$

$$\Rightarrow \qquad 1-\beta \ \ge \ \alpha$$

$$\Rightarrow \qquad P_{\theta_1}(W) \ \ge \ P_{\theta_0}(W)$$

$$\Rightarrow \quad P\,[\mathbf{x} : \mathbf{x} \in W \mid H_1] \ \ge \ P\,[\mathbf{x} : \mathbf{x} \in W \mid H_0] \qquad \dots (16·9a)$$

In other words, the critical region W is said to be unbiased if

$$P_\theta(W) \ \ge \ P_{\theta_0}(W), \forall\, \theta\, (\ne \theta_0) \in \Theta \qquad \dots (16·96b)$$

Theorem 16·2. *Every most powerful (MP) or uniformly most powerful (UMP) critical region (CR) is necessarily unbiased.*

(*i*) *If W be an MPCR of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, then it is necessarily unbiased.*

(*ii*) *Similarly if W be UMPCR of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$, then it is also unbiased.*

Proof. Since W is an MPCR of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, by Neymán-Pearson Lemma, we have ; for $\forall\, k > 0$,

$$W = \{\mathbf{x} : L(\mathbf{x}, \theta_1) \ge k\, L(\mathbf{x}, \theta_0)\} = \{\mathbf{x} : L_1 \ge k\, L_0\}$$

and $W' = \{\mathbf{x} : L(\mathbf{x}, \theta_1) < k\, L(\mathbf{x}, \theta_0)\} = \{\mathbf{x} : L_1 < k\, L_0\}$,

where k is determined so that the size of the test is α *i.e.*,

$P_{\theta_0}(W) = P\,[\mathbf{x} \in W \mid H_0] = \int_W L_0\, d\mathbf{x} = \alpha$ ...(i)

To prove that W is unbiased, we have to show that :

Power of $W \geq \alpha$ *i.e.,* $P_{\theta_1}(W) \geq \alpha$ ...(ii)

We have :

$$P_{\theta_1}(W) = \int_W L_1\, d\mathbf{x} \;\geq\; k \int_W L_0\, d\mathbf{x} = k\alpha$$

[$\because$ On W, $L_1 \geq k\, L_0$ and Using (i)]

i.e., $P_{\theta_1}(W) \geq k\alpha,\ \forall\ k > 0$ ...(iii)

Also

$1 - P_{\theta_1}(W) = 1 - P\,(\mathbf{x} \in W \mid H_1) = P\,(\mathbf{x} \in W' \mid H_1)$

$= \int_{W'} L_1\, d\mathbf{x}$

$< k \int_{W'} L_0\, d\mathbf{x} = k\,P\,(\mathbf{x} : \mathbf{x} \in W' \mid H_0)$

[$\because$ On W', $L_1 < k\, L_0$]

$= k\,[1 - P\,(\mathbf{x} : \mathbf{x} \in W \mid H_0)]$

$= k\,(1 - \alpha)$ [Using (i)]

i.e., $1 - P_{\theta_1}(W) < k\,(1 - \alpha),\ \forall\ k > 0$ ...(iv)

Case (i) $k \geq 1$. If $k \geq 1$, then from (iii), we get

$P_{\theta_1}(W) \geq k\alpha \geq \alpha$

$\Rightarrow$ W is unbiased CR.

Case (ii) $0 < k < 1$. If $0 < k < 1$, then from (iv), we get :

$1 - P_{\theta_1}(W) < 1 - \alpha$

$\Rightarrow$ $P_{\theta_1}(W) > \alpha$

$\Rightarrow$ W is unbiased C.R.

Hence MP critical region is unbiased.

(ii) If W is UMPCR of size α then also the above proof holds if for θ_1 we write θ such that $\theta \in \Theta_1$. So we have

$$P_\theta(W) > \alpha,\ \forall\ \theta \in \Theta_1$$

$\Rightarrow$ W is unbiased CR.

16·5·2. Optimum Regions and Sufficient Statistics. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with p.m.f. or p.d.f. $f(x, \theta)$, where the parameter θ may be a vector. Let T be a sufficient statistic for θ. Then by Factorization Theorem,

$$L\,(\mathbf{x}, \theta) = \prod_{i=1}^{n} f\,(x_i, \theta) = g_\theta\,(t\,(\mathbf{x}))\cdot h(\mathbf{x}) \qquad \ldots(*)$$

where $g_\theta\,(t(\mathbf{x}))$ is the marginal distribution of the statisitc $T = t(\mathbf{x})$.

By Neyman-Pearson Lemma, the MPCR for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is given by :

$$W = \{\mathbf{x} : L(\mathbf{x}, \theta_1) \geq k\, L(\mathbf{x}, \theta_0)\}, \ \forall\ k > 0 \qquad \ldots(**)$$

From (*) and (**), we get

$$W = \{\mathbf{x} : g_{\theta_1}(t(\mathbf{x})) \,.\, h(\mathbf{x}) \geq k .\ g_{\theta_0}(t(\mathbf{x})) \,.\, h(\mathbf{x})\}, \forall\ k > 0$$

$$= \{\mathbf{x} : g_{\theta_1}(t(\mathbf{x})) \geq k \,.\ g_{\theta_0}(t(\mathbf{x}))\}, \forall\ k > 0$$

Hence if $T = t(\mathbf{x})$ is sufficient statistic for θ then the MPCR for the test may be defined in terms of the marginal distribution of $T = t(\mathbf{x})$, rather than the joint distribution of $X_1, X_2, \ldots, X_n$.

Example 16·1. *Given the frequency function :*

$$f(x, \theta) = \frac{1}{\theta},\ 0 \leq x \leq \theta$$
$$= 0, \text{ elsewhere}$$

and that you are testing the null hypothesis $H_0 : \theta = 1$ against $H_1 : \theta = 2$, by means of a single observed value of x. What would be the sizes of the type I and type II errors, if you choose the interval (i) $0{\cdot}5 \leq x$, (ii) $1 \leq x \leq 1{\cdot}5$ as the critical regions ? Also obtain the power function of the test.

[*Gauhati Univ, B.Sc. 1993; Calcutta Univ. B.Sc. (Maths Hons.), 1987*]

Solution. Here we want to test

$H_0 : \theta = 1$, against $H_1 : \theta = 2$.

(*i*) Here $W = \{x : 0{\cdot}5 \leq x\} = \{x : x \geq 0{\cdot}5\}$

and $\bar{W} = \{x : x \leq 0{\cdot}5\}$

$$\alpha = P\{x \in W \mid H_0\} = P\{x \geq 0{\cdot}5 \mid \theta = 1\}$$
$$= P\{0{\cdot}5 \leq x \leq \theta \mid \theta = 1\} = P\{0{\cdot}5 \leq x \leq 1 \mid \theta = 1\}$$
$$= \int_{0\cdot5}^{1} [f(x, \theta)]_{\theta=1}\, dx = \int_{0\cdot5}^{1} 1 .\, dx = 0{\cdot}5$$

Similarly,

$$\beta = P\{x \in \bar{W} \mid H_1\} = P\{x \leq 0{\cdot}5 \mid \theta = 2\}$$
$$= \int_{0}^{0\cdot5} [f(x, \theta)]_{\theta=2}\, dx = \int_{0}^{0\cdot5} \frac{1}{2}\, dx = 0{\cdot}25$$

Thus the sizes of type I and type II errors are respectively

$\alpha = 0{\cdot}5$ and $\beta = 0{\cdot}25$

and power function of the test $= 1 - \beta = 0{\cdot}75$

(*ii*) $W = \{x : 1 \leq x \leq 1{\cdot}5\}$

$$\alpha = P\{x \in W \mid \theta = 1\} = \int_{1}^{1\cdot5} [f(x, \theta)]_{\theta=1}\, dx = 0,$$

since under $H_0 : \theta = 1, f(x, \theta) = 0$, for $1 \leq x \leq 1{\cdot}5$.

$$\beta = P\{x \in \bar{W} \mid \theta = 2\} = 1 - P\{x \in W \mid \theta = 2\}$$

$$= 1 - \int_1^{1.5} [f(x, \theta)]_{\theta=2} dx = 1 - \left| \frac{x}{2} \right|_1^{1.5} = 0.75$$

$\therefore$ Power Function $= 1 - \beta = 1 - 0.75 = 0.25$

Example 16·2. *If $x \geq 1$, is the critical region for testing $H_0 : \theta = 2$ against the alternative $\theta = 1$, on the basis of the single observation from the population,*

$$f(x, \theta) = \theta \exp(-\theta x), 0 \leq x < \infty,$$

obtain the values of type I and type II errors.

[*Poona Univ. M.C.A. 1993; Allahabad Univ. B.Sc., 1993; Delhi Univ. B.Sc (Stat. Hons.), 1988*]

Solution. Here $W = \{x : x \geq 1\}$ and $\bar{W} = \{x : x < 1\}$.

and $H_0 : \theta = 2,\ H_1 : \theta = 1$

$$\begin{aligned}
\alpha &= \text{Size of Type I error} \\
&= P[x \in W | H_0] = P[x \geq 1 | \theta = 2] \\
&= \int_1^{\infty} [f(x, \theta)]_{\theta=2}\, dx \\
&= 2 \int_1^{\infty} e^{-2x}\, dx = 2 \left| \frac{e^{-2x}}{-2} \right|_1^{\infty} \\
&= e^{-2} = 1/e^2 \\
\beta &= \text{Size of type II error} \\
&= P[x \in \bar{W} | H_1] = P\{x < 1 | \theta = 1\} \\
&= \int_0^1 e^{-x}\, dx = \left| \frac{e^{-x}}{-1} \right|_0^1 \\
&= (1 - e^{-1}) = \frac{e - 1}{e}
\end{aligned}$$

Example 16·3. *Let p be the probability that a coin will fall head in a single toss in order to test $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{3}{4}$. The coin is tossed 5 times and H_0 is rejected if more than 3 heads are obtained. Find the probability of type I error and power of the test.*

Solution. Here

$$H_0 : p = \frac{1}{2} \text{ and } H_1 : p = \frac{3}{4}.$$

If the $r.v.\ X$ denotes the number of heads in n tosses of a coin then $X \sim B(n, p)$ so that

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$= \binom{5}{x} p^x (1-p)^{5-x}, \qquad \ldots(*)$$

since $n = 5$, (given). The critical region is given by

$$W = \{x : x \geq 4\} \Rightarrow \overline{W} = \{x : x \leq 3\}$$

$$\begin{aligned}
\alpha &= \text{Probability of type I error} \\
&= P\,[X \geq 4 \mid H_0] \\
&= P[X = 4 \mid p = \tfrac{1}{2}] + P[X = 5 \mid p = \tfrac{1}{2}] \\
&= \binom{5}{4}(\tfrac{1}{2})^4(\tfrac{1}{2})^{5-4} + \binom{5}{5}(\tfrac{1}{2})^5 \qquad \text{[From (*)]} \\
&= 5\,(\tfrac{1}{2})^5 + (\tfrac{1}{2})^5 = 6\,(\tfrac{1}{2})^5 \\
&= \frac{3}{16}
\end{aligned}$$

$$\begin{aligned}
\beta &= \text{Probability of Type II error} \\
&= P[x \in \overline{W} \mid H_1] = 1 - P\,[x \in W \mid H_1] \\
&= 1 - [P(X = 4 \mid p = \tfrac{3}{4}) + P(X = 5 \mid p = \tfrac{3}{4})] \\
&= 1 - \left[\binom{5}{4}(\tfrac{3}{4})^4(\tfrac{1}{4}) + \binom{5}{5}(\tfrac{3}{4})^5\right] \\
&= 1 - (\tfrac{3}{4})^4 \{ \tfrac{5}{4} + \tfrac{3}{4} \} \\
&= 1 - \frac{81}{128} = \frac{47}{128}
\end{aligned}$$

$\therefore$ Power of the test is

$$1 - \beta = \frac{81}{128}$$

Example 16·4. *Let $X \sim N(\mu, 4)$, μ unknown. To test $H_0 : \mu = -1$ against $H_1 : \mu = 1$, based on a sample of size 10 from this population, we use the critical region $x_1 + 2x_2 + \ldots + 10x_{10} \geq 0$. What is its size ? What is the power of the test ?*

Solution. Critical Region $W = \{\mathbf{x} : x_1 + 2x_2 + \ldots + 10x_{10} \geq 0\}$.

Let $U = x_1 + 2x_2 + \ldots + 10x_{10}$

Since x_i's are *i.i.d.* $N(\mu, 4)$,

$$U \sim N\,[(1 + 2 + \ldots + 10)\,\mu,\ (1^2 + 2^2 + \ldots + 10^2)\,\sigma^2] = N\,(55\mu, 385\sigma^2)$$

$$\Rightarrow \quad U \sim N(55\mu, 385 \times 4) = N(55\mu, 1540) \qquad \ldots(*)$$

The size 'α' of the critical region is given by :

$$\alpha = P\,(\mathbf{x} \in W \mid H_0) = P(U \geq 0 \mid H_0) \qquad \ldots(**)$$

Under $H_0 : \mu = -1$, $U \sim N(-55, 1540)$

$$\Rightarrow \qquad Z = \frac{U - E(U)}{\sigma_U} = \frac{U + 55}{\sqrt{1540}}$$

$\therefore$ Under H_0, when $U = 0$, $Z = \dfrac{55}{\sqrt{1540}} = \dfrac{55}{39 \cdot 2428} = 1 \cdot 4015$

$$\therefore \quad \alpha = P\,(Z \geq 1 \cdot 4015) \qquad \text{[From (**)]}$$
$$= 0 \cdot 5 - P(0 \leq Z \leq 1 \cdot 4015)$$
$$= 0 \cdot 5 - 0 \cdot 4192 \qquad \text{(From Normal Probability Tables)}$$
$$= 0 \cdot 0808$$

Alternatively, $\alpha = 1 - P(Z \leq 1 \cdot 4015) = 1 - \Phi(1 \cdot 4015)$,
where $\Phi(\cdot)$ is the distribution function of standard normal variate.

Power of the test is given by :

$$1 - \beta = P(\mathbf{x} \in W \mid H_1) = P(U \geq 0 \mid H_1)$$

Under $H_1 : \mu = 1$, $U \sim N(55, 1540)$

$$\Rightarrow \qquad Z = \frac{U - E(U)}{\sigma_U} = \frac{-55}{\sqrt{1540}} = -1 \cdot 40 \qquad \text{(when } U = 0)$$

$$\therefore \qquad 1 - \beta = P(Z \geq -1 \cdot 40)$$
$$= P(-1 \cdot 4 \leq Z \leq 0) + 0 \cdot 5$$
$$= P(0 \leq Z \leq 1 \cdot 4) + 0 \cdot 5 \qquad \text{(By symmetry)}$$
$$= 0 \cdot 4192 + 0 \cdot 5$$
$$= 0 \cdot 9192$$

Alternatively,

$$1 - \beta = 1 - P(Z \leq -1 \cdot 40) = 1 - \Phi\,(-1 \cdot 40).$$

Example 16·5. *Let X have a p.d.f. of the form :*

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \; ; 0 < x < \infty, \quad \theta > 0$$
$$= 0, \text{ elsewhere.}$$

To test $H_0 : \theta = 2$, against $H_1 : \theta = 1$. use the random sample x_1, x_2 of size 2 and define a critical region :

$$W = \{(x_1, x_2) : 9 \cdot 5 \leq x_1 + x_2\}$$

Find : (i) Power of the test.

(ii) Significance level of the test.

Solution. We are given the critical region :

$$W = \{(x_1, x_2) : 9 \cdot 5 \leq x_1 + x_2\} = \{(x_1\, x_2) : x_1 + x_2 \geq 9 \cdot 5\}$$

Size of the critical region *i.e.*, the significance level of the test is given by :

$$\alpha = P(\mathbf{x} \in W \mid H_0) = P[x_1 + x_2 \geq 9 \cdot 5 \mid H_0] \qquad \ldots(*)$$

In sampling from the given exponential distribution,

$$\frac{2}{\theta} \sum_{i=1}^{n} x_i \sim \chi^2_{(2n)} \qquad \text{[c.f. Example 16·8]}$$

$$\Rightarrow \qquad U = \frac{2}{\theta}(x_1 + x_2) \sim \chi^2_{(4)}, \; (n = 2).$$

$$\therefore \qquad \alpha = P\left[\frac{2}{\theta}(x_1 + x_2) \geq \frac{2}{\theta} \times 9{\cdot}5 \mid H_0\right] \qquad \text{[Form (*)]}$$

$$= P\left[\chi^2_{(4)} \geq 9{\cdot}5\right] \qquad (\because \text{ Under } H_0,\ \theta = 2)$$

$$\Rightarrow \qquad \alpha = 0{\cdot}05 \qquad \text{[From Probability Tables of } \chi^2\text{-distribution]}$$

Power of the test is given by

$$1 - \beta = P(\underset{\sim}{x} \in W \mid H_1) = P(x_1 + x_2 \geq 9{\cdot}5 \mid H_1)$$

$$= P\left[\frac{2}{\theta}(x_1 + x_2) \geq \frac{2}{\theta} \times 9{\cdot}5 \mid H_1\right]$$

$$= P\left[\chi^2_{(4)} \geq 19\right] \qquad (\because \text{ Under } H_1,\ \theta = 1)$$

Example 16·6. *Use the Neyman-Pearson Lemma to obtain the best critical region for testing $\theta = \theta_0$ against $\theta = \theta_1 > \theta_0$ and $\theta = \theta_1 < \theta_0$, in the case of a normal population $N(\theta, \sigma^2)$, where σ^2 is known. Hence find the power of the test.*

[***Delhi Univ. B.Sc. (Stat. Hons), 1986; Gujarat Univ. B.Sc. 1992***]

Solution.

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right]$$

Using Neyman-Pearson Lemma, the best critical region (B.C.R.) is given by (for $k > 0$)

$$\frac{L_1}{L_0} = \frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right]}{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0)^2\right]} \geq k$$

$$\Rightarrow \quad \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i - \theta_1)^2 - \sum_{i=1}^{n}(x_i - \theta_0)^2\right\}\right] \geq k$$

$$\Rightarrow \quad \exp\left[-\frac{n}{2\sigma^2}(\theta_1^2 - \theta_0^2) + \frac{1}{\sigma^2}(\theta_1 - \theta_0)\sum_{i=1}^{n} x_i\right] \geq k$$

$$\Rightarrow \quad -\frac{n}{2\sigma^2}(\theta_1^2 - \theta_0^2) + \frac{1}{\sigma^2}(\theta_1 - \theta_0)\sum_{i=1}^{n} x_i \geq \log k$$

(since $\log x$ is an increasing function of x)

$$\Rightarrow \quad \bar{x}(\theta_1 - \theta_0) \geq \frac{\sigma^2}{n}\log k + \frac{\theta_1^2 - \theta_0^2}{2}$$

Case (*i*) If $\theta_1 > \theta_0$, the B.C.R. is determined by the relation (right-tailed test) :

$$\bar{x} > \frac{\sigma^2}{n} \cdot \frac{\log k}{\theta_1 - \theta_0} + \frac{\theta_1 + \theta_0}{2}$$

i.e., $\bar{x} > \lambda_1$, (say).

$\therefore$ BCR is $W = \{\mathbf{x} : \bar{x} > \lambda_1\}$ ...(16·10)

Case (*ii*) If $\theta_1 < \theta_0$, the B.C.R. is given by the relation (left tailed test)

$$\bar{x} < \frac{\sigma^2}{n} \cdot \frac{\log k}{\theta_1 - \theta_0} + \frac{\theta_1 + \theta_0}{2} = \lambda_2, \text{ (say)}$$

Hence B.C.R. is $W_1 = \{\mathbf{x} : \bar{x} \le \lambda_2\}$ ...(16·11)

The constants λ_1 and λ_2 are so chosen as to make the probability of each of the relations (16·10) and (16·11) equal to α when the hypothesis H_0 is true. The sampling distribution of $\bar{x}$, when H_i is true is $N\left(\theta_i, \frac{\sigma^2}{n}\right)$, $(i = 0, 1)$. Therefore the constants λ_1 and λ_2 are determined from the relations:

$$P[\bar{x} > \lambda_1 \mid H_0] = \alpha \quad \text{and} \quad P[\bar{x} < \lambda_2 \mid H_0] = \alpha$$

$$\therefore \quad P(\bar{x} > \lambda_1 \mid H_0) = P\left[Z > \frac{\lambda_1 - \theta_0}{\sigma/\sqrt{n}}\right] = \alpha \,; Z \sim N(0, 1)$$

$$\Rightarrow \quad \frac{\lambda_1 - \theta_0}{\sigma/\sqrt{n}} = z_\alpha \quad \Rightarrow \quad \lambda_1 = \theta_0 + \frac{\sigma}{\sqrt{n}} z_\alpha \qquad \text{...(16·12)}$$

where z_α is the upper α-point of the standard normal variate given by

$$P(Z > z_\alpha) = \alpha \qquad \text{...(*)}$$

Also $\quad P(\bar{x} < \lambda_2 \mid H_0) = \alpha \Rightarrow P(\bar{x} \ge \lambda_2 \mid H_0) = 1 - \alpha$

$$\Rightarrow \quad P\left(Z \ge \frac{\lambda_2 - \theta_0}{\sigma/\sqrt{n}}\right) = 1 - \alpha \quad \Rightarrow \quad \frac{\lambda_2 - \theta_0}{\sigma/\sqrt{n}} = z_{1-\alpha}$$

$$\Rightarrow \quad \lambda_2 = \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \qquad \text{...(16·12a)}$$

Note. By symmetry of normal distribution, we have $z_{1-\alpha} = -z_\alpha$.

Power of the test. By definition, the power of the test in case (*i*) is :

$$1 - \beta = P[\mathbf{x} \in W \mid H_1] = P[\bar{x} \ge \lambda_1 \mid H_1]$$

$$= P\left[Z \ge \frac{\lambda_1 - \theta_1}{\sigma/\sqrt{n}}\right] \qquad \left[\because \text{ Under } H_1, Z = \frac{\bar{x} - \theta_1}{\sigma/\sqrt{n}} \sim N(0, 1)\right]$$

$$= P\left(Z \ge \frac{\theta_0 + \frac{\sigma}{\sqrt{n}} z_\alpha - \theta_1}{\sigma/\sqrt{n}}\right) \qquad \text{[Using (16·12)]}$$

$$= P\left[Z \ge z_\alpha - \frac{\theta_1 - \theta_0}{\sigma/\sqrt{n}}\right] \qquad (\because \theta_1 > \theta_0)$$

$$= 1 - P(Z \le \lambda_3) \qquad \left[\lambda_3 = z_\alpha - \frac{\theta_1 - \theta_0}{\sigma/\sqrt{n}}, \text{(say).}\right]$$

$$= 1 - \Phi(\lambda_3), \qquad \ldots(16\cdot13)$$

where Φ (.) is the distribution function of standard normal variate.

Similarly in case (*ii*), $(\theta_1 < \theta_0)$, the power of the test is

$$1 - \beta = P(\bar{x} < \lambda_2 \mid H_1) = P\left(Z < \frac{\lambda_2 - \theta_1}{\sigma/\sqrt{n}}\right)$$

$$= P\left[Z < \frac{\theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} - \theta_1}{\sigma/\sqrt{n}}\right] \qquad \text{[Using (16·12a)]}$$

$$= P\left[Z < z_{1-\alpha} + \frac{\theta_0 - \theta_1}{\sigma/\sqrt{n}}\right] \qquad (\because \theta_0 > \theta_1)$$

$$= \Phi(\lambda_4), \qquad \ldots(16\cdot13a)$$

where $$\lambda_4 = z_{1-\alpha} + \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} = \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} - z_\alpha \qquad \ldots(16\cdot13b)$$

UMP Critical Region. (16·10) provides best critical region for testing $H_0 : \theta = \theta_0$ against the hypothesis, $H_1 : \theta = \theta_1$, provided $\theta_1 > \theta_0$ while (16·11) defines the best critical region for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, provided $\theta_1 < \theta_0$. Thus the best critical region for testing simple hypothesis $H_0 : \theta = \theta_0$ against the simple hypothesis, $H_1 : \theta = \theta_1 + c, c > 0$, will not serve as the best critical region for testing simple hypothesis $H_0 : \theta = \theta_0$ against simple alternative hypothesis $H_1 : \theta = \theta_0 - c, c > 0$.

Hence in this problem, no uniformly most powerful test exists for testing the simple hypothesis, $H_0 : \theta = \theta_0$ against the composite alternative hypothesis, $H_1 : \theta \ne \theta_0$.

However, for each alternative hypothesis, $H_1 : \theta = \theta_1 > \theta_0$ or $H_1 : \theta = \theta_1 < \theta_0$, *a UMP* test exists and is given by (16·10) and (16·11) respectively.

Remark. In particular, if we take $n = 2$, then the B.C.R. for testing $H_0 : \theta = \theta_0$, against $H_1 : \theta = \theta_1$ $(> \theta_0)$ is given by : [From (16·10) and (16·12)]

$$W = \{\mathbf{x} : (x_1 + x_2)/2 \ge \theta_0 + \sigma z_\alpha / \sqrt{2}\} \qquad [\because \bar{x} = (x_1 + x_2)/2]$$

$$= \{\mathbf{x} : x_1 + x_2 \ge 2\theta_0 + \sqrt{2}\,\sigma z_\alpha\}$$

$$= \{\mathbf{x} : x_1 + x_2 \ge C\}, \text{ (say)}, \qquad \ldots(**)$$

where $$C = 2\theta_0 + \sqrt{2}\,\sigma z_\alpha = 2\theta_0 + \sqrt{2}\,\sigma \times 1\cdot645, \text{ if } \alpha = 0\cdot05.$$

Similarly, the B.C.R. for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ $(< \theta_0)$ with $n = 2$ and $\alpha = 0\cdot05$ is given by [From (16·11) and (16·12a)] :

$$W_1 = \{\mathbf{x} : (x_1 + x_2)/2 \le \theta_0 - \sigma z_\alpha/\sqrt{2}\}$$
$$= \{\mathbf{x} : (x_1 + x_2) \le 2\theta_0 - \sqrt{2}\,\sigma \times 1{\cdot}645\}$$
$$= \{\mathbf{x} : x_1 + x_2 \le C_1\}\ , \text{(say)}, \qquad \ldots(***)$$

where $\quad C_1 = 2\theta_0 - \sqrt{2}\,\sigma z_\alpha = 2\theta_0 - \sqrt{2}\,\sigma \times 1{\cdot}645.$

The B.C.R. for testing $H_0 : \theta = \theta_0$ against the two tailed alternative $H_1 : \theta = \theta_1\ (\ne \theta_0)$, is given by :

$$W_2 = \{\mathbf{x} : (x_1 + x_2 \ge C) \cup (x_1 + x_2 \le C_1)\} \qquad \ldots(****)$$

The regions in (**), (***), and (****) are given by the shaded portions in the following figures (*i*), (*ii*) and (*iii*) respectively.

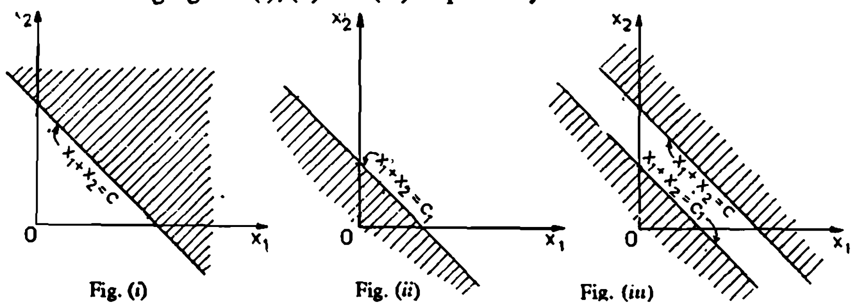


Fig. (*i*) BCR for $\left.\begin{matrix} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1\ (> \theta_0)\end{matrix}\right.$

Fig. (*ii*) BCR for $\left.\begin{matrix} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1\ (< \theta_0)\end{matrix}\right.$

Fig. (*iii*) BCR for $\left.\begin{matrix} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1\ (\ne \theta_0)\end{matrix}\right.$

Example 16·7. *Show that for the normal distribution with zero mean and variance σ^2, the best critical region for $H_0 : \sigma = \sigma_0$ against the alternative $H_1 : \sigma = \sigma_1$ is of the form :*

$$\sum_{i=1}^{n} x_i^2 \le a_\alpha\ , \text{for } \sigma_0 > \sigma_1$$

and
$$\sum_{i=1}^{n} x_i^2 \ge b_\alpha\ , \text{for } \sigma_0 < \sigma_1$$

Show that the power of the best critical region when $\sigma_0 > \sigma_1$ is $F\left(\dfrac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{\alpha, n}\right)$, where $\chi^2_{\alpha,n}$ is lower 100 α-per cent point and $F(\cdot)$ is the distribution function of the χ^2- distribution with n degrees of freedom.

Solution. Here we are given :

$$f(x, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right); -\infty < x < \infty, \sigma > 0.$$

The best critical region (B.C.R.), according to Neyman-Pearson Lemma, is given by (for $k_\alpha > 0$)

$$\frac{L_0}{L_1} \le \frac{1}{k_\alpha} = A_\alpha, \text{(say)}$$

$$\Rightarrow \quad \left(\frac{\sigma_1}{\sigma_0}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)\right\} \le A_\alpha$$

$$\Rightarrow \quad n\log\left(\frac{\sigma_1}{\sigma_0}\right)-\frac{1}{2}\sum_{i=1}^{n} x_i^2\left(\frac{\sigma_1^2-\sigma_0^2}{\sigma_0^2\sigma_1^2}\right) \le \log A_\alpha$$

(since $\log x$ is an increasing function of x).

$$\Rightarrow \quad \frac{\sigma_0^2-\sigma_1^2}{2\sigma_0^2\sigma_1^2}\sum_{i=1}^{n} x_i^2 \le \left[\log A_\alpha - n\log\left(\frac{\sigma_1}{\sigma_0}\right)\right] \qquad \ldots(*)$$

Case (i). If $\sigma_1 < \sigma_0$, then B.C.R. is given by [From (*)]

$$\sum_{i=1}^{n} x_i^2 \le \left[\log A_\alpha - n\log\left(\frac{\sigma_1}{\sigma_0}\right)\right]\frac{2\sigma_0^2\sigma_1^2}{\sigma_0^2-\sigma_1^2} = a_\alpha, \text{ (say)}.$$

i.e., $W = \left\{\mathbf{x} : \sum_{i=1}^{n} x_i^2 \le a_\alpha\right\}$, for $\sigma_1 < \sigma_0$. ...(16·14)

Case (ii). If $\sigma_1 > \sigma_0$, then B.C.R. is given by [From (*)]

$$\sum_{i=1}^{n} x_i^2 \ge \left[\log A_\alpha - n\log\left(\frac{\sigma_1}{\sigma_0}\right)\right]\cdot\frac{2\sigma_0^2\sigma_1^2}{\sigma_0^2-\sigma_1^2} = b_\alpha, \text{ (say)}.$$

i.e., $W_1 = \left\{\mathbf{x} : \sum_{i=1}^{n} x_i^2 \ge b_\alpha\right\}$, for $\sigma_1 > \sigma_0$ ...(16·14a)

The constants a_α and b_α are so chosen that the size of the critical region is α.

Thus a_α is determined so that $P[\mathbf{x} \in W \mid H_0] = \alpha$

$$\Rightarrow \quad P\left[\sum_{i=1}^{n} x_i^2 \le a_\alpha \mid H_0\right] = \alpha$$

$$\Rightarrow \quad P\left[\sum_{i=1}^{n} \frac{x_i^2}{\sigma_0^2} \le \frac{a_\alpha}{\sigma_0^2} \mid H_0\right] = \alpha \qquad \ldots(**)$$

Since under H_0,

$$\chi_{(n)}^2 = \sum_{i=1}^{n}\frac{x_i^2}{\sigma_0^2}, \text{ is a } \chi^2\text{-variate with } n \text{ } d.f.,$$

$$\therefore \quad P\left[\chi_{(n)}^2 \le \frac{a_\alpha}{\sigma_0^2}\right] = \alpha$$

$$\Rightarrow \quad \frac{a_\alpha}{\sigma_0^2} = \chi^2_{\alpha,n} \quad \Rightarrow \quad \sigma_0^2\chi^2_{\alpha,n} = a_\alpha \qquad \ldots(16·15)$$

where $\chi^2_{\alpha,n}$ is the lower 100 α-per cent point of chi-square distribution with n $d.f.$ given by

$$P(\chi^2 \le \chi^2_{\alpha,n}) = \alpha \qquad \ldots(16·15a)$$

Hence the B.C.R. for testing $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma = \sigma_1$ $(< \sigma_0)$, is given by [From (16·14) and (16·15)] :

$$W = \left\{ \mathbf{x} \ : \ \sum_{i=1}^{n} x_i^2 \ \leq \sigma_0^2 \, \chi^2_{\alpha,\, n} \right\} \qquad \ldots(16\cdot15b)$$

where $\chi^2_{\alpha,\, n}$ is defined in (16·15*a*).

Also by definition, the power of the test is :

$$\begin{aligned}
1 - \beta &= P[\mathbf{x} \in W \mid H_1] = P\left[\sum_{i=1}^{n} x_i^2 \leq a_\alpha \mid H_1 \right] \\
&= P\left[\frac{\sum_{i=1}^{n} x_i^2}{\sigma_0^2} \leq \frac{a_\alpha}{\sigma_0^2} \mid H_1 \right] = P\left[\frac{\sum_{i=1}^{n} x_i^2}{\sigma_0^2} \leq \chi^2_{\alpha,\, n} \mid H_1 \right] \\
&= P\left[\frac{\sum_{i=1}^{n} x_i^2}{\sigma_1^2} \leq \frac{\sigma_0^2}{\sigma_1^2} \chi^2_{\alpha,\, n} \mid H_1 \right] \\
&= P\left[\chi^2_{(n)} \leq \frac{\sigma_0^2}{\sigma_1^2} \chi^2_{\alpha,\, n} \right],
\end{aligned}$$

since under H_1, $\sum x_i^2/\sigma_1^2$ is a χ^2-variate with *n d.f.*

Hence, power of the test $= F\left(\dfrac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{\alpha,\, n} \right)$, $\qquad \ldots(16\cdot15c)$

where F(·) is the distribution function of chi-square distribution with n d.f.

Remarks 1. Similarly, for testing $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma = \sigma_1$ $(> \sigma_0)$, b_α in (16·14*a*) is determined so that :

$$P\left[\mathbf{x} \in W_1 \mid H_0\right] = \alpha$$

$$\Rightarrow \quad P\left[\mathbf{x} \ : \ \sum_{i=1}^{n} x_i^2 \geq b_\alpha \mid H_0 \right] = \alpha$$

$$\Rightarrow \quad P\left[\mathbf{x} : \frac{\sum x_i^2}{\sigma_0^2} \geq \frac{b_\alpha}{\sigma_0^2} \mid H_0 \right] = \alpha$$

$$\Rightarrow \quad P\left[\mathbf{x} : \chi^2_{(n)} \geq \frac{b_\alpha}{\sigma_0^2} \right] = \alpha$$

$$\Rightarrow \quad P\left[\mathbf{x} : \chi^2_{(n)} \leq \frac{b_\alpha}{\sigma_0^2} \right] = 1 - \alpha$$

$$\Rightarrow \quad \frac{b_\alpha}{\sigma_0^2} = \chi^2_{1-\alpha,\, n} \quad \Rightarrow \quad b_\alpha = \sigma_0^2 \cdot \chi^2_{1-\alpha,\, n} \qquad \ldots(16\cdot16)$$

where $\chi^2_{\alpha,\, n}$ is defined in (16·15*a*).

Hence the B.C.R. for testing $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma = \sigma_1$ $(> \sigma_0)$, is given by :

$$W_1 = \left\{ \mathbf{x} : \sum_{i=1}^{n} x_i^2 \geq \sigma_0^2 \cdot \chi^2_{1-\alpha,\, n} \right\} \qquad \text{...(16·16}a\text{)}$$

The power of the test in this case is given by

$$1 - \beta = P(\mathbf{x} \in W_1 \mid H_1) = P\left[\sum_{i=1}^{n} x_i^2 \geq \sigma_0^2 \, \chi^2_{1-\alpha,\, n} \mid H_1 \right]$$

$$= P\left[\frac{\sum_{i=1}^{n} x_i^2}{\sigma_1^2} \geq \frac{\sigma_0^2}{\sigma_1^2} \chi^2_{1-\alpha,\, n} \mid H_1 \right]$$

$$= P\left[\chi^2_{(n)} \geq \frac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{1-\alpha,\, n} \right], \qquad \text{...(16·16}b\text{)}$$

since under H_1, $\sum_{i=1}^{n} x_i^2 / \sigma_1^2$ is a χ^2-variate with n d.f.

$$\therefore \quad 1 - \beta = 1 - P\left[\chi^2_{(n)} \leq \frac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{1-\alpha,\, n} \right]$$

$$= 1 - F\left(\frac{\sigma_0^2}{\sigma_1^2} \cdot \chi^2_{1-\alpha,\, n} \right), \qquad \text{...(16·16}c\text{)}$$

where $F(.)$ is the distribution function of chi-square distribution with n d.f.

2. *Graphical representation of the B.C.R. for the particular case $n = 2$.*

For $n = 2$, the B.C.R. for testing $H_0 : \sigma = \sigma_0$, against $H_1 : \sigma = \sigma_1$ ($< \sigma_0$) is given by [From (16·15b)]

$$W = \left\{ \mathbf{x} : \sum_{i=1}^{2} x_i^2 \leq \sigma_0^2 \cdot \chi^2_{\alpha,\, 2} \right\}$$

$$= \left\{ \mathbf{x} : x_1^2 + x_2^2 \leq a^2 \right\},$$

where $a^2 = \sigma_0^2 \chi^2_{\alpha,\, 2}$. Thus the B.C.R. is the interior of the circle with centre (0, 0) and radius 'a' and is shown as the shaded region in Figure (i) on page 16·22.

Similarly, from (16·16a), the B.C.R. for testing $H_0 : \sigma = \sigma_0$, against $H_1 : \sigma = \sigma_1$ ($> \sigma_0$) for $n = 2$ is given by :

$$W_1 = \left\{ \mathbf{x} : x_1^2 + x_2^2 \geq \sigma_0^2 \chi^2_{1-\alpha,\, 2} \right\} = \left\{ \mathbf{x} : x_1^2 + x_2^2 \geq b^2 \right\}$$

where $b^2 = \sigma_0^2 \cdot \chi^2_{1-\alpha,\, 2}$. Thus, B.C.R. is the exterior of the circle with centre (0, 0) and radius b and is shown as the shaded region in Figure (ii) on page 16·22.

Similarly the B.C.R. for testing $H_0 : \sigma = \sigma_0$ against the two-tailed alternative $H_1 : \sigma = \sigma_1$ ($\neq \sigma_0$), for $n = 2$ is given by :

$$W_3 = W_1 \cup W_2$$

$$= \left\{ \mathbf{x} : x_1^2 + x_2^2 \leq a^2 \right\} \cup \left\{ \mathbf{x} : x_1^2 + x_2^2 \geq b^2 \right\}$$

and is shown as the shaded region in the Figure (*iii*) below.

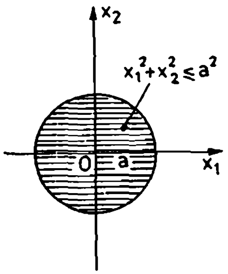


Fig. (*i*)

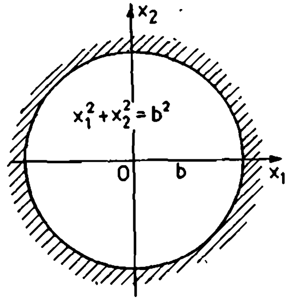


Fig. (*ii*)

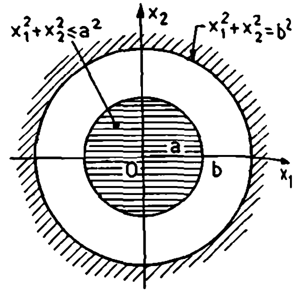


Fig (*iii*)

3. (*16·14*) defines an *UMP* test for testing simple hypothesis $H_0 : \sigma = \sigma_0$ against simple alternative hypothesis $H_1 : \sigma = \sigma_1$ ($< \sigma_0$) whereas (*16·14a*) defines an *UMP* test for testing simple hypothesis $H_0 : \sigma = \sigma_0$ against the simple alternative hypothesis $H_1 : \sigma = \sigma_1$ ($> \sigma_0$). However no *UMP* test exists for testing simple hypothesis $H_0 : \sigma = \sigma_0$ against the composite alternative hypothesis $H_1 : \sigma \ne \sigma_0$.

Example 16·8. *Given a random sample $X_1, X_2, \ldots, X_n$ from the distribution with p.d.f.* $f(x, \theta) = \theta\, e^{-\theta x}, x > 0$
show that there exists no UMP test for testing

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \ne \theta_0.$$

[***Delhi Univ. B.Sc. (Stat. Hons.), 1988; Gorakhpur Univ. B.Sc., 1993***]

Solution. $$L = \prod_{i=1}^{n} f(x_i, \theta) = \theta^n . \exp\left[-\theta \sum_{i=1}^{n} x_i\right]$$

Consider $\qquad H_1 : \theta = \theta_1, (\theta_1 \ne \theta_0)$.

The best critical region, using Neyman-Perason Lemma is given by :

$$\theta_1^n \exp[-\theta_1 \Sigma x_i] \ge k . \theta_0^n \exp[-\theta_0 \Sigma x_i] \; ; k > 0$$

$$\Rightarrow \quad \exp[(\theta_0 - \theta_1) \Sigma x_i] \ge k . \left(\frac{\theta_0}{\theta_1}\right)^n$$

$$\Rightarrow \quad (\theta_0 - \theta_1) \Sigma x_i \ge \log\left[k . \left(\frac{\theta_0}{\theta_1}\right)^n\right] = k_1, \text{ (say)}. \qquad \ldots(*)$$

Case (*i*) If $\theta_1 > \theta_0$, then *B.C.R.* is given by [From (*)]

$$\Sigma x_i \le \frac{k_1}{\theta_0 - \theta_1} = \lambda_1, \text{ (say)}.$$

Case (*ii*) If $\theta_1 < \theta_0$, then *B.C.R.* is given by [From (*)]

$$\Sigma x_i \ge \frac{k_1}{\theta_0 - \theta_1} = \lambda_2, \text{ (say)}.$$

The constants λ_1 and λ_2 are so determined that

$P[\Sigma x_i \le \lambda_1 \mid H_0] = \alpha$ and $P[\Sigma x_i \ge \lambda_2 \mid H_0 = \alpha$

$\Rightarrow P[2\theta \Sigma x_i \le 2\theta\, \lambda_1 \mid H_0] = \alpha$ $\quad\Big|\Rightarrow P[2\theta\, \Sigma x_i \ge 2\theta\, \lambda_2 \mid H_0] = \alpha$

But in random sampling from the given exponential distribution,

$$M_{\Sigma X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left[M_{X_i}((t))\right]^n$$

$$= \left(1 - \frac{t}{\theta}\right)^{-n}$$

$$\Rightarrow \qquad M_{2\theta\Sigma X_i}(t) = M_{\Sigma X_i}(2t\theta) = (1-2t)^{-n},$$

which is the *m.g.f.* of a χ^2-variate with $2n$. *d.f.* Hence by uniqueness theorem of *m.g.f.*'s,

$$2\theta \sum_{i=1}^{n} X_i \sim \chi^2_{(2n)}$$

Using this result in (**)

$$P[2\theta_0 \Sigma x_i \leq \mu_1] = P[\chi^2_{(2n)} \leq \mu_1] = \alpha$$

$$\Rightarrow \qquad \mu_1 = \chi^2_{1-\alpha, 2n}$$

where $\chi^2_{\alpha, n}$ is the upper 'α' point of χ^2-distribution with *n.d.f.* given by

$$P(\chi^2 > \chi^2_{\alpha, n}) = \alpha \qquad \ldots(i)$$

Hence B.C.R. for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ ($> \theta_0$) is given by

$$W_0 = \left\{\mathbf{x} : 2\theta_0 \Sigma x_i \leq \chi^2_{1-\alpha, 2n}\right\}$$

$$= \left\{\mathbf{x} : \Sigma x_i \leq \frac{1}{2\theta_0}\chi^2_{1-\alpha, 2n}\right\}$$

and since it is independent of θ_1, W_0 is *U.M.P.C.R.* for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ ($> \theta_0$).

Similarly from (***), we get

$$P[2\theta_0 \Sigma x_i \geq \mu_2] = P[\chi^2_{(2n)} \geq \mu_2] = \alpha$$

$$\Rightarrow \qquad \mu_2 = \chi^2_{\alpha, 2n}$$

Hence B.C.R. for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ ($< \theta_0$) is given by :

$$W_1 = \left\{\mathbf{x} : 2\theta_0 \Sigma x_i \geq \chi^2_{\alpha, 2n}\right\}$$

$$= \left\{\mathbf{x} : \Sigma x_i \geq \frac{1}{2\theta_0}\chi^2_{\alpha, 2n}\right\},$$

and since it is independent of θ_1, W_1 is also *UPM C.R.* for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ ($< \theta_0$).

However, since the two critical regions W_0 and W_1 are different, there exists no critical region of size α which is *U.M.P.* for $H_0 : \theta = \theta_0$ against the two tailed alternative, $H_1 : \theta \neq \theta_0$.

Power of the test. The power of the test for testing $H_0 : \theta = \theta_0$, against $H_1 : \theta = \theta_1$ ($> \theta_0$) is given by

$$1 - \beta = P[\mathbf{x} \in W_0 \mid H_1]$$

$$= P\left[\sum_{i=1}^{n} x_i \leq \frac{1}{2\theta_0}\chi^2_{1-\alpha, 2n} \mid H_1\right]$$

$$= P\left[2\theta_1 \sum_{i=1}^{n} x_i \le \frac{\theta_1}{\theta_0}\chi^2_{1-\alpha,\, 2n} \mid H_1\right]$$

$$= P\left[\chi^2_{(2n)} \le \frac{\theta_1}{\theta_0}\chi^2_{1-\alpha,\,(2n)}\right], \quad \ldots(*)$$

since under H_1, $2\theta_1 \sum_{i=1}^{n} x_i \sim \chi^2_{(2n)}$.

Similarly the power of the test for testing $H_0 : \theta = \theta_0$, against $H_1 : \theta = \theta_1$ $(< \theta_0)$ is given by :

$$1 - \beta = P[\mathbf{x} \in W_1 \mid H_1]$$

$$= P\left[\sum_{i=1}^{n} x_i \ge \frac{1}{2\theta_0}\chi^2_{\alpha,\, 2n} \mid H_1\right]$$

$$= P\left[2\theta_1 \sum_{i=1}^{n} x_i \ge \frac{\theta_1}{\theta_0}\chi^2_{\alpha,\, 2n} \mid H_1\right]$$

$$= P\left[\chi^2_{(2n)} \ge \frac{\theta_1}{\theta_0}\chi^2_{\alpha,\, 2n}\right] \quad \ldots(**)$$

Remark. The graphic representation of the B.C.R. for $H_0 : \theta = \theta_0$ against different alternatives $H_1 : \theta = \theta_1$ $(> \theta_0)$, $H_1 : \theta = \theta_1$ $(< \theta_0)$ and $H_1 : \theta = \theta_1$ $(\ne \theta_0)$ for $n = 2$, can be done similarly as in Example 16·6, for the mean of normal distribution.

Example 16·9. *For the distribution :*

$$dF = \begin{cases} \beta \exp\{-\beta(x - \gamma)\}\, dx, & x \ge \gamma \\ 0, & x < \gamma \end{cases}$$

show that for a hypothesis H_0 that $\beta = \beta_0$, $\gamma = \gamma_0$ and an alternative H_1 that $\beta = \beta_1$, $\gamma = \gamma_1$, the best critical region is the region given by

$$\bar{x} \le \frac{1}{\beta_1 - \beta_0}\left\{\gamma_1\beta_1 - \gamma_0\beta_0 - \frac{1}{n}\log k + \log\frac{\beta_1}{\beta_0}\right\},$$

provided that the admissible hypothesis is restricted by the condition

$$\gamma_1 \le \gamma_0,\ \beta_1 \ge \beta_0 \qquad \textit{(Gauhati Univ. M.Sc., 1992)}$$

Solution. $f(x ; \beta, \gamma) = \beta \exp\{-\beta(x - \gamma)\},\ x \ge \gamma$
$= 0$, otherwise

$$\therefore \quad \prod_{i=1}^{n} f(x_i ; \beta, \gamma) = \beta^n \exp\{-\beta \sum_{i=1}^{n}(x_i - \gamma)\};\ x_1, x_2, \ldots, x_n \ge \gamma$$

$= 0$, otherwise

Using Neyman-Pearson Lemma, the B.C.R. for $k > 0$, is given by

$$\frac{\beta_1{}^n \exp\left[-\beta_1 \sum_{i=1}^{n} (x_i - \gamma_1)\right]}{\beta_0{}^n \exp\left[-\beta_0 \sum_{i=1}^{n} (x_i - \gamma_0)\right]} \geq k$$

$$\Rightarrow \quad \left(\frac{\beta_1}{\beta_0}\right)^n \exp\left[-\beta_1 \sum_{i=1}^{n} (x_i - \gamma_1) + \beta_0 \sum_{i=1}^{n} (x_i - \gamma_0)\right] \geq k$$

$$\Rightarrow \quad \left(\frac{\beta_1}{\beta_0}\right)^n \exp\left[-\beta_1 n(\bar{x} - \gamma_1) + \beta_0 n(\bar{x} - \gamma_0)\right] \geq k$$

$$\Rightarrow \quad n \log (\beta_1/\beta_0) - n\bar{x}(\beta_1 - \beta_0) + n\beta_1\gamma_1 - n\beta_0\gamma_0 \geq \log k$$

(since $\log x$ is an increasing function of x).

$$\Rightarrow \quad \bar{x}(\beta_1 - \beta_0) \leq \left\{\gamma_1\beta_1 - \gamma_0\beta_0 - \frac{1}{n}\log k + \log\left(\frac{\beta_1}{\beta_0}\right)\right\}$$

$$\Rightarrow \quad \bar{x} \leq \frac{1}{\beta_1 - \beta_0}\left\{\gamma_1\beta_1 - \gamma_0\beta_0 - \frac{1}{n}\log k + \log\left(\frac{\beta_1}{\beta_0}\right)\right\}$$

provided $\beta_1 > \beta_0$.

Example 16·10. *Examine whether a best critical region exists for testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1 > \theta_0$ for the parameter θ of the distribution :*

$$f(x, \theta) = \frac{1 + \theta}{(x + \theta)^2}, \quad 1 \leq x < \infty$$

[*Bangalore Univ. B.Sc., 1992*]

Solution. $\prod_{i=1}^{n} f(x_i, \theta) = (1 + \theta)^n \prod_{i=1}^{n} \frac{1}{(x_i + \theta)^2}$

By Neyman-Pearson Lemma, the B.C.R. is given by

$$(1 + \theta_1)^n \prod_{i=1}^{n} \frac{1}{(x_i + \theta_1)^2} \geq k\,(1 + \theta_0)^n \prod_{i=1}^{n} \frac{1}{(x_i + \theta_0)^2}$$

$$\Rightarrow n \log (1 + \theta_1) - 2 \sum_{i=1}^{n} \log (x_i + \theta_1)$$

$$\geq \log k + n \log (1 + \theta_0) - 2 \sum_{i=1}^{n} \log (x_i + \theta_0)$$

$$\Rightarrow 2 \sum_{i=1}^{n} \log\left(\frac{x_i + \theta_0}{x_i + \theta_1}\right) \geq \log k + n \log\left(\frac{1 + \theta_0}{1 + \theta_1}\right)$$

Thus the test criterion is $\sum_{i=1}^{n} \log \left(\dfrac{x_i + \theta_0}{x_i + \theta_1} \right)$, which cannot be put in the form of a function of the sample observations, not depending on the hypothesis. Hence no B.C.R. exists in this case.

EXERCISE 16(a)

1. (*a*) What are simple and composite statistical hypotheses ? Give examples. Define null and alternative hypotheses. How is a statistical hypothesis tested ?

(*b*) Explain the following terms :

(*i*) Errors of first and second kinds.
(*ii*) The best critical region.
(*iii*) Power function of a test.
(*iv*) Level of significance.
(*v*) Simple and composite hypotheses.
(*vi*) Most powerful test.
(*vii*) Uniformly most powerful test.

(*c*) Identify the composite hypotheses in the following, where μ is the mean and σ^2 is the variance of a distribution.

(*i*) H_0 : $\mu \le 0, \quad \sigma^2 = 1$
(*ii*) H_0 : $\mu = 0, \quad \sigma^2 = 0$
(*iii*) H_0 : $\mu \le 0, \quad \sigma^2 =$ arbitrary
(*iv*) H_0 : $\sigma^2 = \sigma^2_0$ (a given value), μ arbitrary.

(*d*) (*i*) Explain the concepts of Type I and Type II errors, with examples and bring out their importance in Neyman and Pearson testing theory.

2. "In every hypothesis testing, the two types of errors are always present." If this is true then explain what is the use of hypothesis testing.

[*Delhi Univ. M.C.A., 1990*]

3. What is a statistical hypothesis ? Define (*i*) two types of errors, (*ii*) power of a test; with reference to testing of a hypothesis. Explain how the best critical region is determined. State clearly the theorem which is used to determine the best critical region for simple hypothesis at a given significance level.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1992*]

4. Explain the concept of the most powerful tests and discuss how the Neyman-Pearson lemma enables us to obtain the most powerful critical region for testing a simple hypothesis against a simple alternative.

[*Madras Univ. B.Sc., 1988*)

5. What is meant by a statistical hypothesis ? Explain the concepts of type I and type II errors. Show that a most powerful test is necessarily unbiased.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1985*]

6. What are simple and composite statistical hypotheses ? State and prove Neyman-Pearson Fundamental Lemma for testing a simple hypothesis against a simple alternative. *[Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1986]*

7. (*a*) Explain the basic concepts of statistical hypothesis. Discuss the problems associated with the testing of simple and composite hypotheses. Show that a most powerful test is necessarily unbiased.

[Delhi Univ. B.Sc. (Stat. Hons.), 1983]

8. State Neyman-Pearson Lemma.

Prove that if W is an MP region for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, then it is necessarily unbiased. Also prove that the same holds good if W is an UMP region. *[Delhi Univ. B.Sc. (Stat. Hons.), 1982]*

9. (*a*) Let p denote the probability of getting a head when a given coin is tossed once. Suppose that the hypothesis $H_0 : p = 0{\cdot}5$ is rejected in favour of $H_1 : p = 0{\cdot}6$ if 10 trials result in 7 or more heads. Calculate the probabilities of type I and type II errors. *[Calcutta Univ. B.Sc. (Maths Hons.), 1989]*

(*b*) An urn contains 6 marbles of which θ are white and the others black. In order to test the null hypothesis $H_0 : \theta = 3$, against the alternative $H_1 : \theta = 4$, two marbles are drawn at random (without replacement) and H_0 is rejected if both the marbles are white; otherwise H_0 is accepted. Find the probabilities of committing type I and type II errors.

If it is decided to reject H_0 when both marbles are black and to accept it otherwise, find the probabilities of rejecting H_0 (*i*) when H_0 is true and (*ii*) when H_1 is true. Comment on your results.

10. (*a*) p is the probability that a given die shows even number. To test $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{1}{3}$, following procedure is adopted. Toss the die twice and accept H_0 if both times it shows even number. Find the probabilities of type I and type II errors.

(*b*) Let p be the probability that a coin will fall head in a single toss. In order to test the hypothesis $H_0 : p = \frac{1}{2}$, the coin is tossed 6 times and the hypothesis H_0 is rejected if more than 4 heads are obtained. Find the probability of the error of first kind. If the alternative hypothesis is $H_1 : p = \frac{3}{4}$, find the probability of the error of second kind.

(c) In a Bernoulli distribution with parameter p, $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{2}{3}$, is rejected if more than 3 heads are obtained out of 5 throws of a coin. Find the probabilities of Type I and Type II errors.

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

11. (*a*) Let $X_1, X_2, \ldots, X_9$ be a random sample from $N(\theta, 25)$. If, for testing $H_0 : \theta = 20$, against $H_1 : \theta = 26$, the critical region W is defined by

$$W = \{ \mathbf{x} \mid \bar{X} > 23{\cdot}266 \},$$

then find the size of critical region and the power.

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

(*b*) Let $X \sim N(\mu, 4)$, μ unknown. To test $H_0 : \mu = -1$ against $H_1 : \mu = 1$, based on a sample of size 10 from this population we use the critical region $x_1 + 2x_2 + \ldots + 10x_{10} \geq 0$. What is its size ? What is the power of the test ?

(*c*) A sample of size 16 is drawn from a normal population with mean μ and standard deviation σ for testing the hypothesis $H_0 : \mu = \sigma = 1$, against the alternative hypothesis $H_1 : \mu = \sigma = 2$. It is decided to reject the hypothesis H_0 if the sample mean exceeds 1·5 and otherwise accept it.

Calculate the probabilities of errors of the first and second kind in this procedure.

$$\left[\text{Given} \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = 0{\cdot}8413 \text{ and } \int_{-\infty}^{2} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = 0{\cdot}9773\right]$$

12. (*a*) The hypothesis $\mu = 50$, is rejected if the mean of a sample of size 25 is either greater than 70·54 or less than 31·19. Assuming the distribution to be normal with s.d. 50, find the level of significance. Obtain the power function for the test and sketch the power curve with two values above 50 and two values below 50.

(*b*) Calculate the size of the type II error if the type I error is chosen to be $\alpha = 0{\cdot}16$ if you are testing $H_0 : \mu = 7$ against $H_1 : \mu = 6$, for a normal distribution with $\sigma = 2$, by means of a sample of size 25 and if the proper tail of the χ^2 distribution is used as the critical region.

13. (*a*) Given the frequency function :

$$f(x, \theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta$$
$$= 0, \text{ elsewhere}$$

and that you are testing the hypothesis $H_0 : \theta = 1{\cdot}5$ against $H_1 : \theta = 2{\cdot}5$, by means of a single observed value of x, what would be the sizes of the type I and type II errors, if you choose the interval $0{\cdot}8 \leq x$, as the critical region ? Also obtain the power function of the test.

(*b*) It is desired to test the hypothesis $H_0 : \theta = 0$ against $H_1 : \theta > 0$, by observing a random variable X which is uniformly distributed on $[\theta, \theta + 1]$. Given only one observation, sketch the power function of the test whose critical region is defined by $\{x > c\}$. What value of c would you choose ?

Given n observations, derive the general formula of the power function of the test whose critical region is defined by : {at least one x is greater than c} and indicate how you would construct a confidence interval for θ.

14. Let X have a p.d.f. of the form :

$$f(x, \theta) = \frac{1}{\theta} \exp(-x/\theta), \; 0 < x < \infty, \; \theta > 0$$
$$= 0, \text{ elsewhere.}$$

To test $H_0 : \theta = 2$ against $H_1 : \theta = 1$, use a random sample X_1, X_2 of size 2 and define a critical region $C = \{(x_1, x_2) : 9{\cdot}5 \leq x_1 + x_2\}$.

Find (*i*) Power function of the test.

(*ii*) Significance level of the test.

(*b*) Let X have a p.d.f. of the form :

$$f(x\,;\theta) = \theta x^{\theta-1}, 0 < x < 1$$
$$= 0, \text{ elsewhere}$$

To test the simple hypothesis $H_0 : \theta = 1$ against the alternative simple hypothesis $H_1 : \theta = 2$, use a random sample X_1, X_2 of size $n = 2$ and define the critical region to be

$$C = \left\{(x_1, x_2) : \tfrac{3}{4} \leq x_1 x_2\right\}$$

where x_1, x_2 are the values assumed by a sample. Obtain the power function of the test. [*Madras Univ. B.Sc., Stat-Main, 19C1*]

Hint. $Y = -\log X$ has an exponential distribution with parameter θ *i.e.*, $Y \sim \gamma(\theta, 1)$.

15. (*a*) Give a working rule of finding the best critical region for testing a simple hypothesis against a simple alternative.

For a normal (m, σ^2) population with known σ, construct a test for the null hypothesis $H_0 : m = m_0$ against the alternative $m > m_0$.

[*Calcutta Univ. B.Sc., (Maths Hons.), 1989*]

(*b*) Let $(x_1, x_2, ..., x_n)$ be a random sample from $N(\theta, \sigma^2)$, where σ^2 is known. Obtain an *UMP* test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Also find the power function of the test and examine if the test is unbiased.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1986, 1982*]

16. (*a*) Obtain the most powerful test for testing the mean $\mu = \mu_0$ against $\mu = \mu_1$, $(\mu_1 > \mu_0)$ when $\sigma^2 = 1$ in normal population.

(*b*) Obtain the most powerful test of size α for $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ when $\mu_1 > \mu_0$, if the probability density function of the random variable X is

$$f(x, \mu) = \frac{1}{\sqrt{8\pi}} \cdot \exp\left\{-\tfrac{1}{8}(x-\mu)^2\right\}, -\infty < x < \infty$$

17. (*a*) Let $x_1, x_2, ..., x_n$ denote a random sample from the distribution that has p.d.f.

$$f(x, \mu) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\tfrac{1}{2}(x-\mu)^2\right], -\infty < x < \infty$$

It is desired to test $H_0 : \mu = 0$ against $H_1 : \mu = 1$.

(*b*) Let $X_1, X_2, ..., X_n$ denote a random sample from the normal distribution $N(\theta, 1)$, θ is unknown. Show that there is no uniformly most powerful test of the simple hypothesis $H_0 : \theta = \theta_0$, where θ_0 is a fixed number against the alternative composite hypothesis $H_1 : \theta \neq \theta_0$.

18. Let $x_1, x_2, ..., x_n$ be a random sample from $N(\mu, \theta)$, where μ is known. Obtain an UMP test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$. Also find the power function of the test. [*Delhi Univ. B.Sc. (Sat. Hons.), 1985*]

19. Define M.P. region and U.M.P. region. Show that an M.P. region is necessarily unbiased.

Obtain M.P. regions of size α for testing—

(i) $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, $(\theta_1 > \theta_0)$

(ii) $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, $(\theta_1 < \theta_0)$

for $N(\mu, \theta)$, where μ is known.

Show that the tests in (i) and (ii) are U.M.P. against one-sided alternative.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

20. (a) Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal distribution $N(0, \sigma^2)$. Show that there exists a uniformly most powerful test with significance level α for testing

$$H_0 : \sigma^2 = \sigma_1^2 \text{ against } H_1 : \sigma^2 < \sigma_1^2.$$

If $n = 15$, $\alpha = 0.05$ and $\sigma_1^2 = 3$, determine the best critical region and the power function of the above test. *[Gujarat Univ. B.Sc., Oct. 1993]*

(b) State Neyman and Pearson's fundamental lemma and apply it to obtain the test for testing $\sigma^2 = 1$ against $\sigma^2 > 1$, when the sample is from $N(0, \sigma^2)$. Is this test UMP ? Is it unbiased ? Give reasons.

[Indian Civil Services (Main), 1990]

21. A sample of size 25 is drawn from a normal population with unknown mean μ and variance 16. It is required to test the hypothesis $H_0 : \mu = 1.0$ against the alternative $H_1 : \mu = 3.0$ at 5% level of significance, Obtain the most powerful test for testing H_0 against H_1 and state how you will find its power. Is the test uniformly most powerful ?

22. Explain the statistical procedure of testing the following hypothesis regarding the standard deviation (σ) of normal population :

$$H_0 : \sigma = \sigma_0$$
$$H_1 : \sigma = \sigma_1 > \sigma_0$$

Will the test criterion remain the same when σ_1 is changed to $\sigma \neq \sigma_0$?

23. State and prove Neyman-Pearson Lemma. If $x \geq 1$ is the critical region for testing $H_0 : \theta = 2$ against the alternative $H_1 : \theta = 1$, on the basis of a single observation from the population

$$f(x, \theta) = \theta\, e^{-\theta x},\ 0 \leq x < \infty,\ \theta > 0,$$

obtain the values of type I and type II errors and the power function of the test.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(b) Given a random sample $X_1, X_2, \ldots, X_n$ of size n from the distribution with p.d.f.

$$f(x, \theta) = \theta\, e^{-\theta x};\ x > 0,\ 0 < \theta < \infty,$$

show that UMP test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ is given by

$$\left\{ \mathbf{x} \mid \Sigma x_i \geq \frac{1}{2\theta_0} \chi^2_{\alpha, 2n} \right\}$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(c) Explain the Neyman-Pearson Lemma for finding the best critical region for testing a simple hypothesis about the parameter θ of the density function $f(x, \theta)$. Illustrate your answer by constructing the best critical region for testing, $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 < \theta_0$, where θ is the parameter of the distribution with p.d.f.,

$$f(x, \theta) = \theta e^{-\theta x} ; 0 < x < \infty, \theta > 0.$$

[*Meerut Univ. B.Sc., 1993; Poona Univ. B.Sc., Oct. 1991*]

24. (*a*) Two independent observations x_1, x_2 are made on a random variable X with density function :

$$f(x, \theta) = \frac{1}{\theta} \exp(-x/\theta); \ 0 < x < \infty, \ \theta > 0.$$

Test the null hypothesis $H_0 : \theta = 2$ against the alternative $H_1 : \theta = 4$. If H_0 is accepted when $x_1 + x_2 < 9{\cdot}5$, and rejected otherwise, obtain the level of significance and power of the test.

(*b*) Let X_1 be a random sample of size one from a population with p.d.f. $f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}; x \geq 0, \theta > 0$. Obtain : (*i*) the B.C.R. of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ and (*ii*) the power of the test.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1983*]

25. (*a*) Obtain the statistic for testing the hypothesis that the mean of a Poisson population is 2 against the alternative that it is 3, on the basis of n independent observations.

(*b*) Suppose you are testing $H_0 : \lambda = 2$ against $H_1 : \lambda = 1$, where λ is the parameter of the Poisson distribution. Obtain the best critical region of the test.

26. (*a*) Suppose a random sample of size n is taken from the Poisson population $\left(\frac{\exp(-\lambda) . \lambda^x}{x\,!}\right)$, $x = 0, 1, 2, \ldots$. Give the most powerful critical region of size α for testing the hypothesis $\lambda = \lambda_0$ against $\lambda = \lambda_1$, $(\lambda_1 > \lambda_0)$.

How can you use the above result to find a confidence interval for λ ?

Write an expression for the power function of the test for the hypothesis $\lambda = \lambda_0$ against $\lambda > \lambda_0$.

(*b*) $X_1, X_2, \ldots, X_{10}$ is a random sample of size 10 from a Poisson distribution with mean θ. Show that the critical region C defined by $\sum_{i=1}^{10} x_i \geq 3$, is the best critical region for testing $H_0 : \theta = 0{\cdot}1$ against $H_1 : \theta = 0{\cdot}5$.

[*Madras Univ. B.Sc., Oct. 1991*]

27. (*a*) Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution having p.d.f.

$$\begin{aligned} f(x, p) &= p^x (1-p)^{1-x} ; x = 0, 1; 0 < p < 1 \\ &= 0, \text{ elsewhere} \end{aligned}$$

It is desired to test $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{1}{3}$.

(*b*) Suppose X has Bernoulli distribution with probability of success θ. On the basis of a random sample of size n it is proposed to reject the null hypothesis, $H_0 : \theta = \frac{1}{2}$ if

$$(X_1 + X_2 + \ldots + X_n) \geq \frac{3}{8} \text{ or } \leq \frac{5}{8}$$

For $n = 5$, find the level of significance of the test.

28. (a) Let $x_1, x_2, \ldots, x_n$ be a random sample from a Bernoulli distribution with density :

$$f(x ; \theta) = \theta^x (1 - \theta)^{1-x} ; x = 0, 1$$

Obtain a uniformly most powerful size-α test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Would you modify the test if $H_1 : \theta < \theta_0$?

[*Delhi Univ. M.A. (Eco.), 1987*]

(b) The probability that a given machine produces a defective item is p and the quality of the items varies independently from one to another. Given a random sample of $n = 20$ items produced by the machine, what is the form of the best acceptance region for testing $H_0 : p = 0{\cdot}05$ versus $H_1 : p = 0{\cdot}10$? What are the possible values of $\alpha \leq 0{\cdot}1$ (probability of type I error) in this case and the corresponding values of β, the probability of type II error ?

29. Derive a most powerful test of the hypothesis $\theta = \frac{1}{4}$ against the alternative $\theta = \frac{1}{2}$ for the parameter θ in a geometric distribution $\theta (1 - \theta)^x$, $x = 0, 1, 2, \ldots$ based on a random sample of size 2.

30. Describe the method for finding the best critical region of size α for testing a simple hypothesis against simple alternative one. Illustrate it by finding BCR for testing $H_0 : \theta = 0$ against $H_1 : \theta = 1$, for the Cauchy distribution.

$$dF(x) = \frac{dx}{\pi[1 + (x - \theta)^2]}, \quad -\infty < x < \infty$$

based on a random sample of size 1.

31. The distribution of x is :

$$f(x, \theta) = \frac{1}{2}, \theta - 1 \leq x \leq \theta + 1$$
$$= 0, \text{ otherwise}$$

If $H_0 : \theta = 4$ and $H_1 : \theta = 5$, determine the critical region on the right hand tail of the distribution corresponding to $\alpha = 0{\cdot}25$. Also calculate the probability of type II error.

[*Kurukshetra Univ. M.A. (Eco.), 1992*]

32. (a) Define simple and composite hypotheses. State and prove Neyman-Pearson Lemma.

(b) Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from p.d.f.

$$f(x, \theta) = \theta x^{\theta-1}, 0 < x < 1, \theta > 0.$$

Obtain the U.M.P. region of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Also find the power function of the test.

33. (a) Let $x_1, x_2, \ldots, x_n$ be n independent observations on a random variable X with density function

$$f(x, \theta) = \theta x^{\theta - 1} ; 0 < x < 1, \theta > 0$$

Show that the best critical region for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$, can be defined in terms of the geometric mean of $x_1, x_2, \ldots, x_n$.

(*b*) Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f.

$$f(x, \theta) = \begin{cases} \theta x^{\theta - 1}, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

where $0 < \theta < \infty$. Show that the M.P. test of level α for testing $H_0 : \theta = 1$ against the alternative $H_1 : \theta = 2$, is given by the critical region :

$$\left\{ \mathbf{x} \;\middle|\; \prod_{i=1}^{n} x_i > \exp\left[-\tfrac{1}{2} \chi^2_{1-\alpha,\, 2n} \right] \right\}$$

where $\chi^2_{1-\alpha,\, 2n}$ is the lower α - point of the χ^2-distribution with $2n$ d.f.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1987*]

34. Let $X_1, X_2, \ldots, X_n$ be a random sample from a p.d.f.

$$f(x, \theta) = \begin{cases} \theta x^{\theta - 1}, & 0 \le x \le 1,\ \theta > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Find an U.M.P. test of size α for testing $H_0 : \theta = 1$ against $H_1 : \theta > 1$. Also obtain the power function. [*Delhi Univ. B.Sc. (Stat. Hons.), 1992*]

35. Let $X_1, X_2, \ldots, X_n$ be a random sample from discrete distribution with probability function $f(x)$ for which x takes non-negative integral values $0, 1, 2, \ldots$.

According to H_0 :

$$f(x) = \begin{cases} \dfrac{e^{-1}}{x\,!}\ ; & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

According to H_1 :

$$f(x) = \begin{cases} \dfrac{1}{2^{x+1}}\ ; & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

Obtain the critical region of the most powerful test of level α for testing H_0 against H_1. Also find the power of the test for the case $n = 1$ and $k = 1$.

36. H_0 denotes the null hypothesis that a given distribution has the p.d.f.

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},\ -\infty < x < \infty$$

and H_1 denotes the alternative hypothesis that the distribution has the p.d.f.

$$\frac{1}{2} \exp(-|x|),\ -\infty < x < \infty.$$

Obtain the most powerful test for testing H_0 against H_1.

37. It is required to test H_0 against H_1 from a single observation x, where H_0 is the hypothesis that the p.d.f. is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \ (-\infty < x < \infty)$$

and H_1 is the hypothesis that the p.d.f. is

$$f(x) = \frac{2}{\Gamma(1/4)} \exp(-x^4), \ (-\infty < x < \infty)$$

Obtain the most powerful test with level of significance α in this case.

38. State Neyman-Pearson fundamental lemma. With the usual notations, if β is the power of the most powerful test of size α for testing $H_0 : p = p_0$ against $H_1 : p = p_1$, show that $\alpha < \beta$ unless $p_0 = p_1$.

If
$$p_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}, \ -\infty < x < \infty,$$
$$p_1(x) = \frac{1}{\pi} \cdot \frac{1}{1 + (x-\mu)^2}, \ -\infty < x < \infty$$

and μ is known, determine the most powerful test of size α. Calculate its power, if α and μ have specified values.

16·6. Likelihood Ratio Test. Neyman-Pearson Lemma based on the magnitude of the ratio of two probability density functions provides the best test for testing simple hypothesis against simple alternative hypothesis. The best test in any given situation depends on the nature of the population distribution and the form of the alternative hypothesis being considered. In this section we shall discuss a general method of test construction called the *Likelihood Ratio (L.R.) Test* introduced by Neyman and Pearson for testing a hypothesis, simple or composite, against a simple or composite alternative hypothesis. This test is related to the maximum likelihood estimates.

Before defining the test, we give below some notations and terminology.

Parameter Space. Let us consider a random variable X with p.d.f. $f(x, \theta)$. In most common applications, though not always, the functional form of the population distribution is assumed to be known except for the value of some unknown parameter(s) θ which may take any value on a set Θ. This is expressed by writing the p.d.f. in the form $f(x, \theta), \theta \in \Theta$. The set Θ, which is the set of all possible values of θ is called the *parameter space*. Such a situation gives rise not to one probability distribution but a family of probability distributions which we write as $\{f(x, \theta), \theta \in \Theta\}$. For example if $X \sim N(\mu, \sigma^2)$, then the parameter space

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma < \infty\}$$

In particular, for $\sigma^2 = 1$, the family of probability distributions is given by

$$\{N(\mu, 1); \mu \in \Theta\}, \text{ where } \Theta = \{\mu : -\infty < \mu < \infty\}$$

In the following discussion we shall consider a general family of distributions

$$\{f(x : \theta_1, \theta_2, \ldots, \theta_k) : \theta_i \in \Theta, i = 1, 2, \ldots, k\}$$

The null hypothesis H_0 will state that the parameters belong to some subspace Θ_0 of the parameter space Θ.

Let $x_1, x_2, \ldots, x_n$ be a random sample of size $n > 1$ from a population with p.d.f. $f(x, \theta_1, \theta_2, \ldots, \theta_k)$, where Θ, the parameter space is the totality of all points that $(\theta_1, \theta_2, \ldots, \theta_k)$ can assume. We want to test the null hypothesis

$$H_0 : (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta_0$$

against all alternative hypotheses of the type

$$H_1 : (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta - \Theta_0$$

The likelihood function of the sample observations is given by

$$L = \prod_{i=1}^{n} f(x_i ; \theta_1, \theta_2, \ldots, \theta_k) \qquad \ldots(16\cdot16)$$

According to the principle of maximum likelihood, the likelihood equation for estimating any parameter θ_i is given by

$$\frac{\partial L}{\partial \theta_i} = 0, \quad (i = 1, 2, \ldots, k) \qquad \ldots(16\cdot17)$$

Using (16·17), we can obtain the maximum likelihood estimates for the parameters $(\theta_1, \theta_2, \ldots, \theta_k)$ as they are allowed to vary over the parameter space Θ and the subspace Θ_0. Substituting these estimates in (16·16), we obtain the maximum values of the likelihood function for variation of the parameters in Θ and Θ_0 respectively. Then the criterion for the likelihood ratio test is defined as the quotient of these two maxima and is given by

$$\lambda = \lambda(x_1, x_2, \ldots, x_n) = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{\sup_{\theta \in \Theta_0} L(\mathbf{x}, \theta)}{\sup_{\theta \in \Theta} L(\mathbf{x}, \theta)}, \qquad \ldots(16\cdot18)$$

where $L(\hat{\Theta}_0)$ and $L(\hat{\Theta})$ are the maxima of the likelihood function (16·16) with respect to the parameters in the regions Θ_0 and Θ respectively.

The quantity λ is a function of the sample observations only and does not involve parameters. Thus λ being a function of the random variables, is also a random variable. Obvious $\lambda > 0$. Further

$$\Theta_0 \subset \Theta \quad \Rightarrow \quad L(\Theta_0) \le L(\Theta) \quad \Rightarrow \quad \lambda \le 1$$

Hence, we get

$$0 \le \lambda \le 1 \qquad \ldots(16\cdot19)$$

The critical region for testing H_0 (against H_1) is an interval

$$0 < \lambda < \lambda_0, \qquad \ldots(16\cdot20)$$

where λ_0 is some number (< 1) determined by the distribution of λ and the desired probability of type 1 error, *i.e.*, λ_0 is given by the equation :

$$P(\lambda < \lambda_0 \mid H_0) = \alpha \qquad \ldots(16\cdot21)$$

For example, if $g(.)$ is the p.d.f. of λ then λ_0 is determined from the equation :

$$\int_0^{\lambda_0} g(\lambda \mid H_0)\, d\lambda = \alpha \qquad \ldots(16\cdot21a)$$

A test that has critical region defined in (16·20) and (16·21) is a *likelihood ratio test* for testing H_0.

Remark. Equations (16·20) and (16·21) define the critical region for testing the hypothesis H_0 by the likelihood ratio test. Suppose that the distribution of λ is not known but the distribution of some function of λ is known, then this knowledge can be utilized as given in the following theorem.

Theorem 16·3. *If λ is the likelihood ratio for testing a simple hypothesis H_0 and if $U = \phi(\lambda)$ is a monotonic increasing (decreasing) function of λ then the test based on U is equivalent to the likelihood ratio test. The critical region for the test based on U is*

$$\phi(0) < U < \phi(\lambda_0) \quad [\phi(\lambda_0) < U < \phi(0)] \qquad \ldots(16\cdot22)$$

Proof. The critical region for the likelihood ratio test is given by $0 < \lambda < \lambda_0$, where λ_0 is determined by

$$\int_0^{\lambda_0} g(\lambda \mid H_0)\, d\lambda = \alpha \qquad \ldots(*)$$

Let $U = \phi(\lambda)$ be a monotonically increasing function of λ. Then (*) gives

$$\alpha = \int_0^{\lambda_0} g(\lambda \mid H_0)\, d\lambda = \int_{\phi(0)}^{\phi(\lambda_0)} h(u \mid H_0)\, du$$

where $h(u \mid H_0)$ is the p.d.f. of U when H_0 is true. Here the critical region $0 < \lambda < \lambda_0$ transforms to $\phi(0) < U < \phi(\lambda_0)$. However if $U = \phi(\lambda)$ is a monotonic decreasing function of λ, then the inequalities are reversed and we get the critical region as $\phi(\lambda_0) < U < \phi(0)$.

2. If we are testing a simple null hypothesis H_0 then there is a unique distribution determined for λ. But if H_0 is composite, then the distribution of λ may or may not be unique. In such a case the distribution of λ may possibly be different for different parameter points in Θ_0 and then λ_0 is to be chosen such that

$$\int_0^{\lambda_0'} g(\lambda \mid H_0)\, d\lambda \le \alpha \qquad \ldots(16\cdot23)$$

for all values of the parameters in Θ_0.

However, if we are dealing with large samples, a fairly satisfactory situation to this testing of hypothesis problem exists as stated (without proof) in the following theorem.

Theorem 16·4. *Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with p.d.f. $f(x\,;\,\theta_1, \theta_2, \ldots, \theta_k)$ where the parameter space Θ is k-dimensional. Suppose we want to test the composite hypothesis*

$$H_0 : \theta_1 = \theta_1', \theta_2 = \theta_2', \ldots, \theta_r = \theta_r'\,;\, r < k$$

where $\theta_1', \theta_2', \ldots, \theta_r'$ are specified numbers. When H_0 is true, $-2 \log_e \lambda$ is asymptotically distributed as chi-square with r degrees of freedom, i.e., under

H_0, $\qquad -2 \log \lambda \sim \chi_{(r)}^2$, $\quad$ *if n is large.* $\qquad$ ...(16·24)

Since $0 \le \lambda \le 1$, $-2 \log_e \lambda$ is an increasing function of λ and approaches infinity when $\lambda \to 0$, the critical region for $-2 \log \lambda$ being the right hand tail of the chi-square distribution. Thus at the level of significance 'α', the test may be stated as follows :

$$\text{Reject } H_0 \text{ if } -2 \log_e \lambda > \chi_{(r)}^2(\alpha)$$

where $\chi_{(r)}^2(\alpha)$ is the upper α-point of the chi-square distribution with r d.f. given by

$$P[\chi^2 > \chi_{(r)}^2(\alpha)] = \alpha,$$

otherwise H_0 may be accepted.

16·6·1. Properties of Likelihood Ratio Test. Likelihood ratio (*L.R.*) test principle is an intuitive one. If we are testing a simple hypothesis H_0 against a simple alternative hypothesis H_1 then the *LR* principle leads to the samt test as given by the Neyman-Pearson lemma. This suggests that *LR* test has some desirable properties, specially large sample properties.

In *LR* test, the probability of type *I* error is controlled by suitably choosing the cut off point λ_0. *LR* test is generally *UMP* if an *UMP* test at all exists. We state below, the two asymptotic properties of *LR* tests.

1. Under certain conditions, $-2 \log_e \lambda$ has an asymptotic chi-square distribution.

2. Under certain assumptions, *LR* test is consistent.

16·7. In this section we shall illustrate how the likelihood ratio criterion can be used to obtain various standard tests of significance in Statistics.

16·7·1. Test for the Mean of a Normal Population. Let us take the problem of testing if the mean of a normal population has a specified value. Let $(x_1, x_2, \ldots, x_n)$ be a random sample of size n from the normal population with mean μ and variance σ^2, where μ and σ^2 are unknown. Suppose we want to test the (composite) null hypothesis

$$H_0 : \mu = \mu_0 \text{ (specified)}, 0 < \sigma^2 < \infty$$

against the composite alternative hypothesis

$$H_1 : \mu \ne \mu_0 ; 0 < \sigma^2 < \infty$$

In this case the parameter space Θ is given by

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$$

and the subspace Θ determined by the null hypothesis H_0 is given by

$$\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, 0 < \sigma^2 < \infty\}$$

The likelihood function of the sample observations $x_1, x_2, \ldots, x_n$ is given by

$$L = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] \qquad \text{...(16·25)}$$

The maximum likelihood estimates of μ and σ^2 are given by :

$$\left. \begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2 \end{aligned} \right\} \qquad \text{...(16·26)}$$

Hence substituting in (16·25), the maximum of L in the parameter space Θ is given by

$$L(\hat{\Theta}) = \left[\frac{1}{2\pi s^2} \right]^{n/2} \cdot \exp\left(-\frac{n}{2} \right) \qquad \text{...(16·27)}$$

In Θ_0, the only variable parameter is σ^2 and *MLE* of σ^2 for given $\mu = \mu_0$ is given by

$$\hat{\sigma}^2 = \frac{1}{n} \Sigma(x_i - \mu_0)^2 = s_0^2, \text{ (say)} \qquad \text{..(16·28)}$$

$$= \frac{1}{n} \Sigma(x_i - \bar{x} + \bar{x} - \mu_0)^2$$

$$= \frac{1}{n} \Sigma(x_i - \bar{x})^2 + (\bar{x} - \mu_0)^2,$$

the product term vanishes, since

$$\Sigma(x_i - \bar{x})(\bar{x} - \mu_0) = (\bar{x} - \mu_0)\, \Sigma(x_i - \bar{x}) = 0$$

$$\therefore \qquad \hat{\sigma}^2 = s^2 + (\bar{x} - \mu_0)^2 = s_0^2, \text{ (say)}. \qquad \text{...(16·28}a\text{)}$$

Hence substituting in (16·25), we get

$$L(\hat{\Theta}_0) = \left[\frac{1}{2\pi\, s_0^2} \right]^{n/2} \exp(-n/2) \qquad \text{...(16·28}b\text{)}$$

The ratio of (16·28*b*) and (16·27) gives the likelihood ratio criterion

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left[\frac{s^2}{s_0^2} \right]^{n/2} \qquad \text{...(16·29)}$$

$$= \left[\frac{s^2}{s^2 + (\bar{x} - \mu_0)^2} \right]^{n/2} = \left\{ \frac{1}{1 + [(\bar{x} - \mu_0)^2/s^2]} \right\}^{n/2} \qquad \text{...(16·29}a\text{)}$$

We have proved earlier (§ 14·2) that under H_0, the statistic

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

where $$S^2 = \frac{1}{n-1} \Sigma(x_i - \bar{x})^2 = \frac{ns^2}{n-1},$$

follows Student's t-distribution with $(n-1)$ d.f.

Thus

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} \sim t_{n-1} \qquad \ldots(16{\cdot}30)$$

Substituting in (16·29a), we get

$$\lambda = \frac{1}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}} = \phi(t^2), \text{ (say).} \qquad \ldots(16{\cdot}31)$$

The likelihood ratio test for testing H_0 against H_1 consists in finding a critical region of the type $0 < \lambda < \lambda_0$, where λ_0 is given by (16·21a), which requires the distribution of λ under H_0. In this case, it is not necessary to obtain the distribution of λ since $\lambda = \phi(t^2)$ is a monotonic function of t^2 and the test can well be carried on with t^2 as a criterion as with λ [c.f. Theorem 16·1]. Now $t^2 = 0$ when $\lambda = 1$ and t^2 becomes infinite when $\lambda = 0$. The critical region of the LR test *viz.*, $0 < \lambda < \lambda_0$, on using (16·31) is equivalent to

$$\left(1 + \frac{t^2}{n-1}\right)^{-n/2} \leq \lambda_0$$

$$\Rightarrow \quad \left(1 + \frac{t^2}{n-1}\right)^{n/2} \geq \lambda_0^{-1}$$

$$\Rightarrow \quad \frac{t^2}{n-1} \geq (\lambda_0)^{-2/n} - 1$$

$$\Rightarrow \quad t^2 \geq (n-1)\left[\lambda_0^{-2/n} - 1\right] = A^2, \text{ (say).}$$

Thus the critical region may well be defined by

$$|t| = \left| \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{S} \right| \geq A \qquad \ldots(16{\cdot}32)$$

where the constant A is determined such that

$$P[|t| \geq A \mid H_0] = \alpha \qquad \ldots(16{\cdot}33)$$

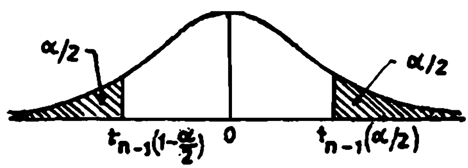
Since under H_0, the statistic t follows Student's t distribution with $(n-1)$ d.f.,

$$A = t_{n-1}(\alpha/2)$$

where the symbol $t_n(\alpha)$ stands for the *right tail* 100 α% point of the t-distribution with n d.f. given by

$$P\{t > t_n(\alpha)\} = \int_{t_n(\alpha)}^{\infty} f(t)\, dt = \alpha \qquad \text{...(16·33a)}$$

where $f(\cdot)$ is the p.d.f. of Student's t with n d.f. The critical region is shown in the following diagram.



Thus for testing $H_0 : \mu = \mu_0$ against $\mu \neq \mu_0$ (σ^2-unknown), we have the two-tailed t-test defined as follows :

If $|t| = \left|\dfrac{\sqrt{n}\,(\bar{x} - \mu_0)}{S}\right| > t_{n-1}(\alpha/2)$, *reject H_0 and if $|t| < t_{n-1}(\alpha/2)$, H_0 may be accepted.*

Important Remarks. 1. Let us now consider the problem of testing the hypothesis

$$H_0 : \mu = \mu_0,\ 0 < \sigma^2 < \infty$$

against the alternative hypothesis

$$H_1 : \mu > \mu_0,\ 0 < \sigma^2 < \infty$$

Here $\quad \Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty,\ 0 < \sigma^2 < \infty\}$

and $\quad \Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0,\ 0 < \sigma^2 < \infty\}$

The maximum likelihood estimates of μ and σ^2 belonging to Θ are given by

$$\hat{\mu} = \begin{cases} \bar{x}, & \text{if } \bar{x} \geq \mu_0 \\ \mu_0, & \text{if } \bar{x} < \mu_0 \end{cases} \qquad \text{...(16·34)}$$

and

$$\hat{\sigma}^2 = \begin{cases} s^2, & \text{if } \bar{x} \geq \mu_0 \\ s_0^2, & \text{if } \bar{x} < \mu_0 \end{cases} \qquad \text{...(16·34a)}$$

where

$$s_0^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2 \qquad \text{...(16·34b)}$$

Thus

$$\hat{L(\Theta)} = \begin{cases} \left(\dfrac{1}{2\pi s^2}\right)^{n/2} \exp\left(-\dfrac{n}{2}\right), & \text{if } \bar{x} \ge \mu_0 \\ \left(\dfrac{1}{2\pi s_0^2}\right)^{n/2} .\exp\left(-\dfrac{n}{2}\right), & \text{if } \bar{x} < \mu_0 \end{cases} \qquad \ldots(16\cdot35)$$

In Θ_0, the only unknown parameter is σ^2 whose *MLE* is given by $\hat{\sigma}^2 = s_0^2$. Thus

$$L(\hat{\Theta}_0) = \left(\frac{1}{2\pi s_0^2}\right)^{n/2} \exp\left(-\frac{n}{2}\right) \qquad \ldots(16\cdot36)$$

$$\therefore \qquad \lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \begin{cases} (s^2/s_0^2)^{n/2}, & \text{if } \bar{x} \ge \mu_0 \\ 1 & \text{if } \bar{x} < \mu_0 \end{cases} \qquad \ldots(16\cdot37)$$

Thus the sample observations $(x_1, x_2, \ldots, x_n)$ for which $\bar{x} < \mu_0$ are to be included in the acceptance region. Hence for the sample observations for which $\bar{x} \ge \mu_0$, the likelihood ratio criterion becomes

$$\lambda = (s^2/s_0^2)^{n/2}, \ \bar{x} \ge \mu_0 \qquad \ldots(16\cdot37a)$$

which is the same as the expression obtained in (16·29). Proceeding similarly as in the above problem, the critical region of the form $0 < \lambda < \lambda_0$ will be equivalently given by [c.f. (16·32)]

$$t^2 = \frac{n(\bar{x} - \mu_0)^2}{S^2} \ge A^2$$

or by

$$t = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{S} \ge A \quad (\because \bar{x} \ge \mu_0) \qquad \ldots(16\cdot38)$$

where t follows Student's t distribution with $(n-1)$ d.f. The constant A is to be determined so that

$$P(t > A) = \alpha \qquad \ldots(16\cdot39)$$

$$\Rightarrow \qquad A = t_{n-1}(\alpha)$$

Hence for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, we have the right tailed-t-test defined as follows :

$$\textit{Reject } H_0 \textit{ if } t = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{S} > t_{n-1}(\alpha) \textit{ and}$$

if $\qquad t < t_{n-1}(\alpha)$, *$H_0$ may be accepted.*

2. If we want to test

$$H_0 : \mu = \mu_0,\ 0 < \sigma^2 < \infty$$

against the alternative hypothesis

$$H_1 : \mu < \mu_0, 0 < \sigma^2 < \infty,$$

then proceeding exactly similarly as in Remark 1 above, we shall get the critical region given by

$$t < -t_{n-1}(\alpha) \quad \text{...(16·40)}$$

In this case we have the left tailed t-test defined as follows :

If $t = \dfrac{\sqrt{n}\,(\bar{x} - \mu_0)}{S} < -t_{n-1}(\alpha)$, *reject* H_0 *otherwise* H_0 *may be accepted.*

3. We summarise below in a tabular form the test criterion, along with the confidence interval for the parameter for testing the hypothesis $H_0 : \mu = \mu_0$ against various alternatives for *the normal population when* σ^2 is not known.

[Here $t_n(\alpha)$ is upper α-point of the t-distrbution with n d.f. as defined in (16·33a).]

NORMAL POPULATION $N(\mu, \sigma^2)$; σ^2 UNKNOWN

Serial No.	*Hypothesis*	*Test*	*Test Statistic*	*Reject H_0 at Level of Significance, α, if*	*$(1-\alpha)$ confidence interval for μ*
1.	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	Two tailed test	$t = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}}$	$\|t\| > t_{n-1}(\alpha/2)$	$\bar{x} - \dfrac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \le \mu$ $\le \bar{x} + \dfrac{S}{\sqrt{n}} t_{n-1}(\alpha/2)$
2.	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	Right tailed test	—do—	$t > t_{n-1}(\alpha)$	$\mu \ge \bar{x} - \dfrac{S}{\sqrt{n}} t_{n-1}(\alpha)$
3.	$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	Left tailed test	— do —	$t < -t_{n-1}(\alpha)$	$\mu \le \bar{x} + \dfrac{S}{\sqrt{n}} t_{n-1}(\alpha)$

16·7·2. Test for the Equality of Means of Two Normal Populations. Let us consider two independent random variables X_1 and X_2 following normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively where the means μ_1, μ_2 and the variances σ_1^2, σ_2^2 are unspecified. Suppose we want to test the hypothesis :

$$H_0 : \mu_1 = \mu_2 = \mu, \text{ (say), (unspecified)}; \; 0 < \sigma_1^2 < \infty, 0 < \sigma_2^2 < \infty,$$

against the alternative hypothesis

$$H_1 : \mu_1 \neq \mu_2, \sigma_1^2 > 0, \sigma_2^2 > 0.$$

Case I. Population variances are unequal.

$$\Theta = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty, \sigma_i^2 > 0, i = 1, 2\}$$

and
$$\Theta_0 = \{(\mu, \sigma_1^2, \sigma_2^2) : -\infty < \mu < \infty, \sigma_i^2 > 0, i = 1, 2\}$$

Let x_{1i} $(i = 1, 2, \ldots, m)$ and x_{2j} $(j = 1, 2, \ldots, n)$ be two independent random samples of sizes m and n from the populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. Then the likelihood function is given by

$$L = \left(\frac{1}{2\pi\sigma_1^2}\right)^{m/2} . \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i} - \mu_1)^2\right]$$
$$\times \left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2} . \exp\left[-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n}(x_{2j} - \mu_2)^2\right] \qquad \ldots(16{\cdot}41)$$

The maximum likelihood estimates for μ_1, μ_2, σ_1^2 and σ_2^2 are given by the equations :

$$\left.\begin{aligned}
&\frac{\partial}{\partial\mu_1}\log L = 0 \Rightarrow \hat{\mu}_1 = \frac{1}{m}\sum_{i=1}^{m} x_{1i} = \bar{x}_1\\
&\frac{\partial}{\partial\mu_2}\log L = 0 \Rightarrow \hat{\mu}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{2j} = \bar{x}_2\\
&\frac{\partial}{\partial\sigma_1^2}\log L = 0 \Rightarrow \hat{\sigma}_1^2 = \frac{1}{m}\sum_{i=1}^{m}(x_{1i} - \bar{x}_1)^2 = s_1^2, \text{ (say).}\\
\text{and } &\frac{\partial}{\partial\sigma_2^2}\log L = 0 \Rightarrow \hat{\sigma}_2^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2 = s_2^2, \text{ (say).}
\end{aligned}\right\} \ldots(16{\cdot}41a)$$

Substituting in (16·41), we get

$$L(\hat{\Theta}) = \left(\frac{1}{2\pi s_1^2}\right)^{m/2} . \left(\frac{1}{2\pi s_2^2}\right)^{n/2} . e^{-(m+n)/2} \qquad \ldots(16{\cdot}42)$$

In Θ_0, we have $\mu_1 = \mu_2 = \mu$ and the likelihood function is given by :

$$L(\Theta_0) = \left(\frac{1}{2\pi\sigma_1^2}\right)^{m/2} . \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i} - \mu)^2\right]$$
$$\times \left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2} . \exp\left[-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n}(x_{2j} - \mu)^2\right]$$

To obtain the maximum value of $L(\Theta_0)$ for variations in μ, σ_1^2 and σ_2^2, it will be seen that estimate of μ is obtained as the root of a cubic equation

$$\frac{m^2(\bar{x}_1 - \mu)}{\sum_{i=1}^{m}(x_{1i} - \hat{\mu})^2} + \frac{n^2(\bar{x}_2 - \mu)}{\sum_{j=1}^{n}(x_{2j} - \hat{\mu})^2} \qquad \ldots(16{\cdot}43)$$

and is thus a complicated function of the sample observations. Consequently the likelihood ratio criterion λ will be a complex function of the observations and its distribution is quite tedious since it involves the ratio of two variances. Consequently, it is impossible to obtain the critical region $0 < \lambda < \lambda_0$, for given α, since the distribution of the population variances is ordinarily unknown. However, in any given instance the cubic equation (16·43) can be solved for μ by numerical analysis technique and thus λ can be computed. Finally, as an approximate test, $-2 \log_e \lambda$ can be regarded as a χ^2-variate with 1 d.f. (c.f. Theorem 16·2).

Case 2. Population Variances are equal, *i.e.*, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, (say). In this case

$$\Theta = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu_i < \infty, \sigma^2 > 0, (i = 1, 2)\}$$

$$\Theta_0 = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

The likelihood function is then given by

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{(m+n)/2} . \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{m}(x_{1i} - \mu_1)^2 + \sum_{j=1}^{n}(x_{2j} - \mu_2)^2\right\}\right] \quad ...(16·44)$$

For $\mu_1, \mu_2, \sigma^2 \in \Theta$, the maximum likelihood equations are given by

$$\left.\begin{aligned} \frac{\partial}{\partial\mu_1}\log L = 0 \;&\Rightarrow\; \hat{\mu}_1 = \bar{x}_1 \\ \frac{\partial}{\partial\mu_2}\log L = 0 \;&\Rightarrow\; \hat{\mu}_2 = \bar{x}_2 \end{aligned}\right\} \quad ...(16·45)$$

and $\dfrac{\partial}{\partial\sigma^2}\log L = 0 \;\Rightarrow\; \hat{\sigma}^2 = \dfrac{1}{m+n}\left[\Sigma(x_{1i} - \hat{\mu}_1)^2 + \Sigma(x_{2j} - \hat{\mu}_2)^2\right]$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{1}{m+n}\left[\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2j} - \bar{x}_2)^2\right]$$

$$= \frac{1}{m+n}\left[m s_1^2 + n s_2^2\right] \quad ...(16·45a)$$

Substituting the values from (16·45) and (16·45*a*) in (16·44), we get

$$L(\hat{\Theta}) = \left[\frac{(m+n)}{2\pi(m s_1^2 + n s_2^2)}\right]^{(m+n)/2} . \exp\left[-\tfrac{1}{2}(m+n)\right] \quad ...(16·46)$$

In Θ_0, $\mu_1 = \mu_2 = \mu$, (say), and we get

$$L(\Theta_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{(m+n)/2} . \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{m}(x_{1i} - \mu)^2 + \sum_{j=1}^{n}(x_{2j} - \mu)^2\right\}\right] \quad ...(16·47)$$

$$\Rightarrow \log L(\Theta_0) = C - \frac{m+n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_i(x_{1i} - \mu)^2 + \sum_j(x_{2j} - \mu)^2\right],$$

where C is a constant independent of μ and σ^2. The likelihood equation for estimating μ gives

$$\frac{\partial}{\partial \mu} \log L = \frac{1}{\sigma^2}\left[\sum_{i=1}^{m} (x_{1i} - \mu) + \sum_{j=1}^{n} (x_{2j} - \mu)\right] = 0$$

$$\Rightarrow \qquad (m\bar{x}_1 + n\bar{x}_2) - (m+n)\mu = 0$$

$$\Rightarrow \qquad \hat{\mu} = \frac{1}{m+n}\left[m\bar{x}_1 + n\bar{x}_2\right] \qquad \text{...(16·48)}$$

Also $\qquad \dfrac{\partial}{\partial \sigma^2} \log L = 0$

$$\Rightarrow \qquad -\frac{(m+n)}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\Sigma(x_{1i} - \mu)^2 + \Sigma(x_{2j} - \mu)^2\right] = 0$$

$$\Rightarrow \qquad \hat{\sigma}^2 = \frac{1}{m+n}\left[\Sigma(x_{1i} - \hat{\mu})^2 + \Sigma(x_{2j} - \hat{\mu})^2\right] \qquad \text{...(16·49)}$$

But $\qquad \displaystyle\sum_{i=1}^{m} (x_{1i} - \hat{\mu})^2 = \sum_{i=1}^{m} (x_{1i} - \bar{x}_1 + \bar{x}_1 - \hat{\mu})^2$

$$= \Sigma(x_{1i} - \bar{x}_1)^2 + m(\bar{x}_1 - \hat{\mu})^2,$$

the product term vanishes since

$$\sum_i (x_{1i} - \bar{x}_1) = 0$$

$$\therefore \qquad \sum_{i=1}^{m} (x_{1i} - \hat{\mu})^2 = ms_1^2 + m\left[\bar{x}_1 - \frac{m\bar{x}_1 + n\bar{x}_2}{m+n}\right]^2$$

$$= ms_1^2 + \frac{mn^2(\bar{x}_1 - \bar{x}_2)^2}{(m+n)^2}$$

Similarly, we shall get

$$\sum_{j=1}^{n} (x_{2j} - \hat{\mu})^2 = ns_2^2 + \frac{nm^2(\bar{x}_2 - \bar{x}_1)^2}{(m+n)^2}$$

Substituting in (16·49), we get

$$\hat{\sigma}^2 = \frac{1}{m+n}\left[ms_1^2 + ns_2^2 + \frac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)^2\right] \qquad \text{...(16·49}a\text{)}$$

Substituting from (16·48) and (16·49a) in (16·47), we get

$$L(\hat{\Theta}_0) = \left\{\frac{(m+n)}{2\pi\left(ms_1^2 + ns_2^2 + \dfrac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)^2\right)}\right\}^{(m+n)/2}$$

$$\times \exp\left(-\frac{m+n}{2}\right) \qquad \text{...(16·50)}$$

$$\therefore \qquad \lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left\{\frac{ms_1^2 + ns_2^2}{ms_1^2 + ns_2^2 + \dfrac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)^2}\right\}^{(m+n)/2}$$

$$= \left\{ \frac{1}{\left(1 + \frac{mn\,(\bar{x}_1 - \bar{x}_2)^2}{(m+n)(ms_1^2 + ns_2^2)}\right)} \right\}^{(m+n)/2} \quad ...(16{\cdot}51)$$

We know that (c.f. § 14·2·10), under the null hypothesis $H_0 : \mu_1 = \mu_2$, the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \quad ...(16{\cdot}52)$$

where
$$S^2 = \frac{1}{m+n-2}(ms_1^2 + ns_2^2) \quad ...(16{\cdot}52a)$$

follows Student's t-distribution with $(m + n - 2)$ d.f. Thus in terms of t, we get

$$\lambda = \left[1 + \frac{t^2}{m+n-2}\right]^{-(m+n)/2} \quad ...(16{\cdot}53)$$

As in § 16·7·1, the test can as well be carried with t rather than with λ. The critical region $0 < \lambda < \lambda_0$, transforms to the critical region of the type

$$t^2 > (m + n - 2)\left[\frac{1}{\lambda_0^{2/(m+n)}} - 1\right] = A^2, \text{ (say)}$$

i.e., by $\quad |t| > A$,

where A is determined so that

$$P\,[\,|t| > A \mid H_0] = \alpha \quad ...(16{\cdot}55)$$

Since under H_0, the statistic t follows Student's t-distribution with $(m + n - 2)$ d.f., we get from (16·55)

$$A = t_{m+n-2}\,(\alpha/2) \quad ...(16{\cdot}56)$$

where, $t_n(\alpha)$ is the right 100 α% point of the t-distribution with n d.f.

Thus for testing the null hypothesis

$$H_0 : \mu_1 = \mu_2\,;\ \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$$

against the alternative

$$H_1 : \mu_1 \neq \mu_2,\ \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0,$$

we have the two-tailed t-test defined as follows ;

If
$$|t| = \left| \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \right| > t_{m+n-2}\,(\alpha/2)$$

reject H_0, otherwise H_0 may be accepted.

Remarks. 1. Proceeding similarly as in Remarks to § 16·7·1, we can obtain the critical regions for testing

$$H_0 : \mu_1 = \mu_2\,;\ \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$$

against the alternative hypothesis

$$H_1 : \mu_1 > \mu_2 \,;\, \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$$

or $$H_1' : \mu_1 < \mu_2 \,;\, \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$$

We give below, in a tabular form the critical region, the test statistic and the confidence interval for testing the hypothesis

$$H_0 : \delta = \mu_1 - \mu_2 = \delta_0, \text{ (say)},$$

against various alternatives, *viz.*, $\delta > \delta_0$, $\delta < \delta_0$ or $\delta \neq \delta_0$.

2. For testing $H_0 : \delta = \delta_0$ against the alternative $H_1 : \delta < \delta_0$, the roles of x_1 and x_2 are interchanged and the case 1 of the table is applied.

3. If $\delta_0 = 0$, the above test reduces to testing $H_0 : \mu_1 = \mu_2$, *i.e.*, the equality of two population means.

4. If the two population variances are not equal, then for testing $H_0 : \delta = \delta_0$, we use Fisher-Behrens' d-test.

S. No.	*Alternative Hypothesis*	*Test*	*Test statistic*	*Reject H_0 at level of significance α if*	*$(1-\alpha)$ confidence interval of δ*
1.	$\delta > \delta_0$	Right tailed	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$	$t > t_{m+n-2}(\alpha) = t_1$, (say)	$\delta \geq (\bar{x}_1 - \bar{x}_2) - t_1 S \sqrt{\frac{1}{m} + \frac{1}{n}}$
2.	$\delta \neq \delta_0$	Two tailed	—do—	$\lvert t \rvert > t_{m+n-2}(\alpha/2) = t_2$, (say)	$(\bar{x}_1 - \bar{x}_2) - t_2 S \sqrt{\frac{1}{m} + \frac{1}{n}} \leq \delta \leq (\bar{x}_1 - \bar{x}_2) + t_2 S \sqrt{\frac{1}{m} + \frac{1}{n}}$

16·7·3. Test for the Equality of Means of Several Normal Populations. Let X_{ij}, $(j = 1, 2, \ldots, n_i;\ i = 1, 2, \ldots k)$ be k independent random samples from k normal populations with means $\mu_1, \mu_2, \ldots, \mu_k$ respectively and unknown but common variance σ^2. In other words, the k normal populations are supposed to be *homoscedastic*. We want to test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu \text{ (say), (unspecified)}$$
$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2 \text{ (say), (unspecified)}$$

against the alternative hypothesis

$$H_1 : \mu_i\text{'s are not all equal,}$$
$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2, \text{ (unspecified)}$$

Thus we have

$$\Theta = \{(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) : -\infty < \mu_i < \infty, (i = 1, 2, \ldots, k) ; \sigma^2 > 0\}$$

and $\Theta_0 = \{(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) : -\infty < \mu_i = \mu < \infty, (i = 1, 2, \ldots, k) : \sigma^2 > 0\}$

The likelihood function of the sample observations is given by

$$L(\Theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} . \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2\right] \quad \ldots(16\cdot57)$$

where $n = \sum_{i=1}^{k} n_i$.

For variations of μ_i, $(i = 1, 2, \ldots, k)$ and σ^2 in Θ, the maximum likelihood estimates are given by

$$\frac{\partial}{\partial \mu_i} \log L(\Theta) = 0 \quad \Rightarrow \quad \sum_j (x_{ij} - \mu_i) = 0$$

$$\Rightarrow \quad \hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} = \bar{x}_i \quad \ldots(16\cdot58)$$

$$\frac{\partial}{\partial \sigma^2} \log L(\Theta) = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum_i \sum_j (x_{ij} - \hat{\mu}_i)^2$$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = \frac{S_W}{n}, \text{ (say)}, \quad \ldots(16\cdot58a)$$

where in *ANOVA* (Analysis of Variance) terminology, S_W is called ***within sample sum of squares (W.S.S.)***.

In Θ_0, the only variable parameters are μ and σ^2 and we have

$$L(\Theta_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} . \exp\left\{-\frac{1}{2\sigma^2} \sum_i \sum_j (x_{ij} - \mu)^2\right\} \quad \ldots(16\cdot59)$$

The *MLE's* of μ and σ^2 are given by

$$\frac{\partial}{\partial \mu} \log L(\Theta_0) = 0 \quad \Rightarrow \quad \sum_i \sum_j (x_{ij} - \mu) = 0$$

$$\Rightarrow \quad \hat{\mu} = \frac{1}{n} \sum\sum x_{ij} = \bar{x} \quad \ldots(16\cdot60)$$

and

$$\frac{\partial}{\partial \sigma^2} \log L(\Theta_0) = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum\sum (x_{ij} - \hat{\mu})^2$$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum\sum (x_{ij} - \bar{x})^2 = \frac{S_T}{n}, \text{ (say)}, \quad \ldots(16\cdot60a)$$

where in *ANOVA* terminology, S_T, is called *total sum of squares (T.S.S.)*

Substituting from (16·58) and (16·58*a*) in (16·57) and from (16·60) and (16·60*a*) in (16·59), we get respectively

$$L(\hat{\Theta}) = \left(\frac{n}{2\pi S_W}\right)^{n/2} . \exp\left(-\frac{n}{2}\right) \qquad \ldots(16{\cdot}61)$$

and
$$L(\hat{\Theta}_0) = \left(\frac{n}{2\pi\ S_T}\right)^{n/2} . \exp\left(-\frac{n}{2}\right) \qquad \ldots(16{\cdot}62)$$

$$\therefore \quad \lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left(\frac{S_W}{S_T}\right)^{n/2} \qquad \ldots(16{\cdot}63)$$

We have

$$S_T = \sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2$$

$$= \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i \sum_j (\bar{x}_i - \bar{x})^2 + 2\sum_i \left[(\bar{x}_i - \bar{x}) \sum_j (x_{ij} - \bar{x}_i)\right]$$

But $\sum\limits_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = 0$, being the algebraic sum of the deviations of the observations of the ith sample from its mean.

$$\therefore \quad S_T = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i n_i(\bar{x}_i - \bar{x})^2$$

$$= S_W + S_B, \text{ (say)} \qquad \ldots(16{\cdot}63a)$$

where $S_B = \sum\limits_i n_i(\bar{x}_i - \bar{x})^2$, in *ANOVA* terminology is called *between samples sum of squares (B.S.S.).*

Substituting in (16·63), we get

$$\lambda = \left(\frac{S_W}{S_W + S_B}\right)^{n/2} \qquad \ldots(16{\cdot}64)$$

$$= \frac{1}{\left[1 + \dfrac{S_B}{S_W}\right]^{n/2}}$$

We know that under H_0, the statistic

$$F = \frac{S_B/(k-1)}{S_W/(n-k)} \qquad \ldots(16{\cdot}65)$$

follows F-distribution with $(k-1; n-k)$ d.f.

Substituting in (16·64), the likelihood ratio criterion λ in terms of F is given by

$$\lambda = \left[1 + \frac{k-1}{n-k} F\right]^{-n/2} \qquad \ldots(16{\cdot}66)$$

Since λ is a monotonic function of F, the test can well be carried on with F as test statistic rather than with λ. The critical region for testing H_0 against H_1, *viz.*, $0 < \lambda < \lambda_0$, is equivalently given by

$$\left[1 + \frac{k-1}{n-k} F\right]^{n/2} > \lambda_0^{-1}$$

$$\Rightarrow \qquad F > \frac{n-k}{k-1}\left[(\lambda_0)^{-2/n} - 1\right] = A, \text{ (say)}, \qquad \text{...(16·67)}$$

where A is determined from the equation

$$P[F > A \mid H_0] = \alpha \qquad \text{...(16·67a)}$$

Since F follows F-distribution with $(k-1, n-k)$ d.f., we get

$$A = F_{k-1,\, n-k}(\alpha)$$

where $F_{k-1,\, n-k}(\alpha)$ denotes the upper α-point of the F-distribution with $(k-1, n-k)$ d.f.

Hence the test for testing

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu,\ \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2 > 0$$

against the alternative hypothesis

$$H_1 : \mu_i\text{'s are not all equal},\ \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2 > 0$$

is defined as follows :

Reject H_0 if $F > F_{k-1,\, n-k}(\alpha)$, otherwise H_0 may be accepted, where F is defined in (16·65).

Remark. In *ANOVA* terminology, $S_B/(k-1)$ is called Between Samples Mean Sum of Squares (M.S.S.) while $S_W/(n-k)$ is called Within Samples (or Error) Mean Sum of Squares and thus F is defined as

$$F = \frac{\text{Between Samples M.S.S.}}{\text{Within Samples M.S.S.}} \qquad \text{...(16·67c)}$$

16·7·4. Test for the Variance of a Normal Population. Let us now consider the problem of testing if the variance of a normal population has a specified value σ_0^2, on the basis of a random sample $x_1, x_2, \ldots, x_n$ of size n from normal population $N(\mu, \sigma^2)$.

We want to test the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2, \text{ (specified)},$$

against the alternative hypothesis

$$H_1 : \sigma^2 \neq \sigma_0^2$$

Here we have

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma^2 > 0\}$$

and

$$\Theta_0 = \{(\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma^2 = \sigma_0^2\}$$

The likelihood function of the sample observations is given by

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} . \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right\} \qquad \text{... (16·68)}$$

As in § (16·7·1), [c.f. (16·27)], we shall get

$$L(\hat{\Theta}) = \left(\frac{1}{2\pi s^2}\right)^{n/2} \exp\left(-\frac{n}{2}\right) \qquad \text{...(16·69)}$$

In Θ_0, we have only one variable parameter, *viz.*, μ and

$$L(\Theta_0) = \left(\frac{1}{2\pi\sigma_0^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \mu)^2\right] \qquad \text{...(16·70)}$$

The *MLE* for μ is given by

$$\frac{\partial}{\partial\mu} \log L = 0 \quad \Rightarrow \quad \hat{\mu} = \bar{x}$$

$$\therefore \qquad L(\hat{\Theta}_0) = \left(\frac{1}{2\pi\sigma_0^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \bar{x})^2\right]$$

$$= \left(\frac{1}{2\pi\sigma_0^2}\right)^{n/2} \exp\left[-\frac{ns^2}{2\sigma_0^2}\right] \qquad \text{...(16·71)}$$

The likelihood ratio criterion is given by

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left[\frac{s^2}{\sigma_0^2}\right]^{n/2} \exp\left[-\frac{1}{2}\left(\frac{ns^2}{\sigma_0^2} - n\right)\right]$$

We know that under H_0, the statistic

$$\chi^2 = \frac{ns^2}{\sigma_0^2} \qquad \text{...(16·72)}$$

follows chi-square distribution with $(n-1)$ d.f. In terms of χ^2, we have

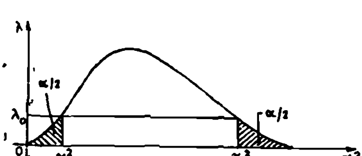
$$\lambda = \left[\frac{\chi^2}{n}\right]^{n/2} . \exp\left[-\frac{1}{2}(\chi^2 - n)\right] \qquad \text{...(16·73)}$$

Since λ is a monotonic function of χ^2, the test may be done using χ^2 as a criterion. The critical region $0 < \lambda < \lambda_0$ is now equivalent to

$$(\chi^2/n)^{n/2} \exp\left[-\frac{1}{2}(\chi^2 - n)\right] < \lambda_0$$

or
$$\exp\left(-\tfrac{1}{2}\chi^2\right)(\chi^2)^{n/2} < \lambda_0 (ne^{-1})^{n/2} = B, \text{ (say).} \qquad \text{...(16·74)}$$

Since χ^2 has chi-square distribution with $(n-1)$ d.f., the critical region (16·74) is determined by a pair of intervals $0 < \chi^2 < \chi_2^2$ and $\chi_1^2 < \chi^2 < \infty$, where χ_1^2 and χ_2^2 are to be determined such that the ordinates of (16·73) are equal, *i.e.*,



$$(\chi_1^2)^{n/2} \exp\left(-\tfrac{1}{2}\chi_1^2\right) = (\chi_2^2)^{n/2} \exp\left(-\tfrac{1}{2}\chi_2^2\right)$$

Critical region is shown as shaded region in the above diagram.

In other words, χ_1^2 and χ_2^2 are defined by the equations

$$\left.\begin{aligned} P(\chi^2 > \chi_1^2) &= \alpha/2 \\ \text{and} \quad P(\chi^2 > \chi_2^2) &= 1 - \frac{\alpha}{2} \end{aligned}\right\} \qquad \ldots(16\cdot75)$$

In other words,

$$\chi_1^2 = \chi^2_{n-1}(\alpha/2) \text{ and } \chi_2^2 = \chi^2_{n-1}(1 - \alpha/2),$$

where $\chi^2_{(n-1)}(\alpha)$ is the upper α-point of the chi-square distribution with $(n-1)$ d.f. Thus the *critical region for testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$, is a two-tailed region given by*

$$\chi^2 > \chi^2_{n-1}(\alpha/2) \quad \text{and} \quad \chi^2 < \chi^2_{n-1}(1 - \alpha/2) \qquad \ldots(16\cdot76)$$

Thus, in this case we have a two-tailed test.

Remarks. 1. If we want to test $H_0 : \sigma^2 = \sigma_0^2$ against the alternative hypothesis $H_1 : \sigma^2 < \sigma_0^2$ we get a one-tailed (left-tailed) test with critical region $\chi^2 < \chi^2_{(n-1)}(1 - \alpha)$ while for testing H_0 against $H_1 : \sigma^2 > \sigma_0^2$, we have a right tailed test with critical region $\chi^2 > \chi^2_{(n-1)}(\alpha)$.

We give below in a tabular form, the test statistic, the test criterion and the confidence interval for the parameter for testing $H_0 : \sigma^2 = \sigma_0^2$, μ (unknown), against various alternative hypotheses.

NORMAL POPULATION $N(\mu, \sigma^2)$; μ UNKNOWN ; $H_0 : \sigma^2 = \sigma_0^2$

S. No.	*Alternative Hypothesis*	*Test*	*Test statistic*	*Reject H_0 at 'α' level of significance if*	*$(1-\alpha)$ confidence interval for σ^2*
1.	$\sigma^2 > \sigma_0^2$	Right-tailed test	$\chi^2 = \dfrac{ns^2}{\sigma_0^2}$	$\chi^2 > \chi^2_{n-1}(\alpha)$	$\sigma^2 \geq \dfrac{ns^2}{\chi^2_{n-1}(\alpha)}$
2.	$\sigma^2 < \sigma_0^2$	Left-tailed test	—do—	$\chi^2 < \chi^2_{n-1}(1-\alpha)$	$\sigma^2 \leq \dfrac{ns^2}{\chi^2_{n-1}(1-\alpha)}$
3.	$\sigma^2 \neq \sigma_0^2$	Two-tailed test	—do—	$\chi^2 > \chi^2_{n-1}(\alpha/2)$ and $\chi^2 < \chi^2_{n-1}(1-\alpha/2)$	$\dfrac{ns^2}{\chi^2_{n-1}(\alpha/2)} \leq \sigma^2 \leq \dfrac{ns^2}{\chi^2_{n-1}(1-\alpha/2)}$

2. If we want to test the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ against the various alternative hypotheses, *viz.*, $\sigma^2 > \sigma_0^2$ or $\sigma^2 < \sigma_0^2$ or $\sigma^2 \neq \sigma_0^2$ for the normal population $N(\mu, \sigma^2)$, where μ is known then the test statistic, the critical region and the confidence interval for σ^2 can be obtained from the table given above on replacing $(n-1)$ by n and ns^2 by the expression $\sum\limits_{i=1}^{n} (x_i - \mu)^2$.

16·7·5. Test for Equality of Variances of two Normal Populations. Consider two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ where the means μ_1 and μ_2 and variances σ_1^2, σ_2^2 are unspecified. We want to test the hypothesis :

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ (unspecified), with μ_1 and μ_2 (unspecified) against the alternative hypothesis

$H_1 : \sigma_1^2 \neq \sigma_2^2$; μ_1 and μ_2 (unspecified).

If x_{1i}, $(i = 1, 2, \ldots, m)$ and x_{2j}, $(j = 1, 2, \ldots, n)$ be independent random samples of sizes m and n from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively then

$$L = \left(\frac{1}{2\pi\sigma_1^2}\right)^{m/2} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i}-\mu_1)^2\right]$$

$$\times \left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n}(x_{2j}-\mu_2)^2\right] \qquad \ldots(16\cdot77)$$

In this case

$$\Theta = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty;\ \sigma_i^2 > 0,\ (i = 1, 2)\}$$

and

$$\Theta_0 = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu_i < \infty;\ (i = 1, 2),\ \sigma^2 > 0\}$$

As in § 16·7·2 [c.f. (16·42)],

$$L(\hat{\Theta}) = \left(\frac{1}{2\pi s_1^2}\right)^{m/2} \cdot \left(\frac{1}{2\pi s_2^2}\right)^{n/2} \cdot \exp\left[-\frac{1}{2}(m+n)\right] \qquad \ldots(16\cdot78)$$

where s_1^2 and s_2^2 are as defined in (16·41a).

In Θ_0, the likelihood function (16·77) is given by

$$L(\Theta_0) = \left[\frac{1}{2\pi\sigma^2}\right]^{(m+n)/2} \cdot \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_i (x_{1i}-\mu_1)^2 + \sum_j (x_{2j}-\mu_2)^2\right\}\right] \qquad \ldots(16\cdot79)$$

and the MLE's for μ_1, μ_2 and σ^2 are now given by

$$\hat{\mu}_1 = \bar{x}_1, \quad \hat{\mu}_2 = \bar{x}_2 \qquad \ldots(16\cdot80)$$

and
$$\hat{\sigma}^2 = \frac{1}{(m+n)}\left[\sum_i (x_{1i}-\hat{\mu}_1)^2 + \sum_j (x_{2j}-\hat{\mu}_2)^2\right]$$

$$= \frac{1}{m+n}\left[\sum_i (x_{1i}-\bar{x}_1)^2 + \sum_j (x_{2j}-\bar{x}_2)^2\right]$$

$$= \frac{ms_1^2 + ns_2^2}{m+n} \qquad \ldots(16\cdot80a)$$

Substituting from (16·80) and (16·80a) in (16·79), we get

$$L(\hat{\Theta}_0) = \left[\frac{m+n}{2\pi(ms_1^2 + ns_2^2)}\right]^{(m+n)/2} \cdot \exp\left[-\frac{1}{2}(m+n)\right] \qquad \ldots(16\cdot81)$$

$$\therefore \quad \lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})}$$

$$= (m+n)^{(m+n)/2} \left\{ \frac{(s_1^2)^{m/2} (s_2^2)^{n/2}}{[ms_1^2 + ns_2^2]^{(m+n)/2}} \right\}$$

$$= \frac{(m+n)^{(m+n)/2}}{m^{m/2} \cdot n^{n/2}} \left\{ \frac{(ms_1^2)^{m/2} (ns_2^2)^{n/2}}{[ms_1^2 + ns_2^2]^{(m+n)/2}} \right\} \qquad \ldots(16{\cdot}82)$$

We know that under H_0, the statistic

$$F = \frac{\Sigma(x_{1i} - \bar{x}_1)^2/(m-1)}{\Sigma(x_{2j} - \bar{x}_2)^2/(n-1)} = \frac{S_1^2}{S_2^2}, \qquad \ldots(16{\cdot}83)$$

follows F-distribution with $(m-1, n-1)$ d.f. (16·83) also implies

$$F = \frac{m(n-1)\, s_1^2}{n(m-1) s_2^2}$$

$$\Rightarrow \quad \left(\frac{m-1}{n-1}\right) F = \frac{ms_1^2}{ns_2^2} \qquad \ldots(16{\cdot}83a)$$

Substituting in (16·82) and simplifying, we get

$$\lambda = \frac{(m+n)^{(m+n)/2}}{m^{m/2}\, n^{n/2}} \left\{ \frac{\left(\frac{m-1}{n-1} F\right)^{m/2}}{\left[1 + \frac{m-1}{n-1} F\right]^{(m+n)/2}} \right\} \qquad \ldots(16{\cdot}84)$$

Thus λ is a monotonic function of F and hence the test can be carried on with F, defined in (16·83) as test statistic. The critical region $0 < \lambda < \lambda_0$ can be equivalently seen to be given by pair of intervals $F \le F_1$ and $F \ge F_2$, where F_1 and F_2 are determined so that under H_0

$$P(F \ge F_2) = \alpha/2 \text{ and } P(F \ge F_1) = 1 - \alpha/2$$

Since, under H_0, F follows Snedecor's F-distribution with $(m-1, n-1)$ d.f., we have

$$F_2 = F_{m-1, n-1}\,(\alpha/2) \text{ and } F_1 = F_{m-1, n-1}\,(1 - \alpha/2),$$

where $F_{m,\,n}(\alpha)$ is the upper α-point of F-distribution with (m, n) d.f. Consequently for testing $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $H_1 : \sigma_1^2 \ne \sigma_2^2$, we have a two-tailed F-test, the *critical region* being given by

$$F > F_{m-1, n-1}\,(\alpha/2) \text{ and } F < F_{m-1, n-1}\,(1 - \alpha/2) \qquad \ldots(16{\cdot}85)$$

where F is defined in (16·83) or (16·83a).

Remark. Let us suppose that we want to test the hypothesis

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = \delta_0^2$$

Without loss of generality, we can assume that $S_1^2 > S_2^2$, where S_1^2 and S_2^2 are unbiased estimates of σ_1^2 and σ_2^2 respectively. We know that the statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \cdot \frac{1}{\delta_0^2}, \text{ (under } H_0\text{)},$$

follows F-distribution with $(m - 1, n - 1)$ d.f. The test-statistic, the test criterion and $(1 - \alpha)$ confidence interval for the parameter for various alternative hypotheses are given in the following table.

If $\delta_0 = 1$, the above test reduces to testing the equality of population variances.

NORMAL POPULATION; $H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = \delta_0^2$

S. No.	Alternative Hypothesis	Test	Test Statistic	Critical region at level of significance 'α'	$(1 - \alpha)$ confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$
1.	$\dfrac{\sigma_1^2}{\sigma_2^2} > \delta_0^2$	Right-tailed	$F = \dfrac{S_1^2}{S_2^2} \cdot \dfrac{1}{\delta_0^2}$	$F > F_{m-1, n-1}(\alpha)$	$\dfrac{\sigma_1^2}{\sigma_2^2} \geq \dfrac{S_1^2}{S_2^2} \times \dfrac{1}{F_{m-1,n-1}(\alpha)}$
2.	$\dfrac{\sigma_1^2}{\sigma_2^2} < \delta_0^2$	Left-tailed	—do—	$F < F_{m-1, n-1}(1 - \alpha)$	$\dfrac{\sigma_1^2}{\sigma_2^2} \leq \dfrac{S_1^2}{S_2^2} \times \dfrac{1}{F_{m-1,n-1}(1-\alpha)}$
3.	$\dfrac{\sigma_1^2}{\sigma_2^2} \neq \delta_0^2$	Two-tailed	—do—	$F > F_{m-1, n-1}(\alpha/2)$, and $F < F_{m-1, n-1}(1 - \alpha/2)$	$\dfrac{S_1^2}{S_2^2} \cdot \dfrac{1}{F_{m-1,n-1}(\alpha/2)} \leq \dfrac{\sigma_1^2}{\sigma_2^2} \leq \dfrac{S_1^2}{S_2^2} \cdot \dfrac{1}{F_{m-1,n-1}(1-\alpha/2)}$

16·7·6. Test for the Equality of Variances of Several Normal Populations. Let X_{ij}, $(j = 1, 2, \ldots, n_i)$ be a random sample of size n_i from the normal population $N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, k$. We want to test the null hypothesis :

$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$ (unspecified), with $\mu_1, \mu_2, \ldots, \mu_k$ (unspecified), against the alternative hypothesis :

$H_1 : \sigma_i^2$ $(i, 2, \ldots, k)$, are not all equal; $\mu_1, \mu_2, \ldots, \mu_k$ (unspecified).

Here we have

$$\Theta = \{(\mu_1, \mu_2, \ldots, \mu_k ; \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) : -\infty < \mu_i < \infty, \sigma_i^2 > 0, \quad (i = 1, 2, \ldots, k)\}$$

$$\text{and } \Theta_0 = \{(\mu_1, \mu_2, \ldots, \mu_k ; \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) : -\infty < \mu_i < \infty, \sigma_i^2 = \sigma^2 > 0, \quad (i = 1, 2, \ldots, k)\}$$

The likelihood function of the sample observations x_{ij}, $(j = 1, 2, \ldots, n_i ;$ $i = 1, 2, \ldots, k)$ is given by :

$$L = \prod_{i=1}^{k} \left\{ \left(\frac{1}{2\pi\, \sigma_i^2} \right)^{n_i/2} . \exp\left[- \frac{1}{2\sigma_i^2} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 \right] \right\} \qquad ...(16\cdot86)$$

It can be easily seen that in Θ the MLE's of μ_i's and σ_i's are given by

$$\hat{\mu}_i = \bar{x}_i \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = s_i^2 \qquad ...(16\cdot87)$$

$$\therefore \qquad L(\hat{\Theta}) = \prod_{i=1}^{k} \left\{ \left(\frac{1}{2\pi s_i^2} \right)^{n_i/2} . \exp\left(- \frac{n_i}{2} \right) \right\}$$

$$= \exp\left(- \frac{n}{2} \right) . \prod_{i=1}^{k} \left[\left(\frac{1}{2\pi s_i^2} \right)^{n_i/2} \right] \qquad ...(16\cdot88)$$

where $n = \sum n_i$.

In Θ_0, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$ and therefore

$$L(\Theta_0) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left[- \frac{1}{2\sigma^2} \sum_j \sum_i (x_{ij} - \mu_i)^2 \right] \qquad ...(16\cdot89)$$

The *MLE*'s of μ_i's and σ^2 are given by

$$\hat{\mu}_i = \bar{x}_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = \frac{1}{n} \sum_i n_i s_i^2 \qquad ...(16\cdot90)$$

Substituting from (16·90) in (16·89), we get

$$L(\hat{\Theta}_0) = \left(\frac{n}{2\pi \sum_i n_i s_i^2} \right)^{n/2} . \exp\left(- \frac{n}{2} \right) \qquad ...(16\cdot91)$$

$$\therefore \qquad \lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{n^{n/2} \prod_{i=1}^{k} \left[(s_i^2)^{n_i/2} \right]}{\left[\sum_{i=1}^{k} n_i s_i^2 \right]^{n/2}}$$

$$= \frac{\prod_{i=1}^{k} \left[(s_i^2)^{n_i/2} \right]}{(s^2)^{\sum n_i/2}}, \text{ where } s^2 = \frac{1}{n} \sum n_i s_i^2$$

$$= \prod_{i=1}^{k} \left[\left(\frac{s_i^2}{s^2} \right)^{n_i/2} \right] \qquad ...(16\cdot92)$$

λ is thus a complicated function of sample observations and it is not easy to obtain its distribution. However, if n_i's are large ($i = 1, 2, \ldots, k$), Theorem 16·2 provides an approximate test defined as follows:

For large n_i's, the quantity $-2 \log_e \lambda$ is approximately distributed as a chi-square variate with $2k - (k + 1) = k - 1$ d.f.

The test can, however, be made even if n_i's are not large. It has been investigated and found that the distribution of $- 2 \log_e \lambda$ is approximately a

χ^2-distribution with $(k-1)$ d.f. even for small n_i's. However, a better approximation is provided by the Bartlett's test statistic :

$$\chi^2 = \frac{-2 \log \lambda'}{1 + \frac{1}{3(k-1)} \left[\sum_i \left(\frac{1}{n_i} \right) - \frac{1}{\sum_i n_i} \right]}$$

where λ' is obtained from λ on replacing n_i by $(n_i - 1)$ in (16·92), which follows χ^2-distribution with $(k-1)$ d.f. Thus the test statistic, under H_0 is given by

$$\chi^2 = \frac{\sum_{i=1}^{k} (n_i - 1) \log_e \left(\frac{s^2}{s_i^2} \right)}{1 + \frac{1}{3(k-1)} \left[\sum_i \left(\frac{1}{n_i} \right) - \frac{1}{\sum_i n_i} \right]} \sim \chi^2_{k-1} \qquad \ldots(16·93)$$

The critical region for the test is, of course, the right-tail of the χ^2-distribution given by

$$\chi^2 > \chi^2_{(k-1)}(\alpha), \qquad \ldots(16·94)$$

where χ^2 is defined in (16·93).

EXERCISE 16 (b)

1. (*a*) Define 'Likelihood Ratio Test'. Under what circumstances would you recommend this test ?

(*b*) Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal distribution $N(\theta_1, \theta_2)$. Use likelihood ratio test to obtain *BCR* of size α under $H_0 : \theta_1 = 0$ against $H_1 : \theta_1 \neq 0$.

2. (*a*) Let $p_\theta(x)$ be the density of a random variable with the mixed second derivative $\frac{\partial^2 \log p_\theta(x)}{\partial\theta\, \partial x} \geq 0$ for all x and θ. Then show that the family has monotone likelihood ratio in x.

3. Discuss the general method of construction of likelihood ratio test. Consider n Bernoullian trials with probability of success p for each trial. Derive the likelihood-ratio test for testing $H_0 : p = p_0$ against $H_1 : p > p_0$.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986*]

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter θ. Derive the likelihood ratio test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Show that this is identical with the corresponding *UMP* test.

5. (*a*) Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with unknown mean μ and, known variance σ^2. Develop the likelihood ratio test for testing $H_0 : \mu = \mu_0$ (specified) against (*i*) $H_1 : \mu > \mu_0$ and (*ii*) $H_1 : \mu < \mu_0$.

(*b*) Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where σ^2 is known. Develop the likelihood ratio test for testing $H_0 : \mu = \mu_0$ (specified) against $H_1 : \mu < \mu_0$. [*Delhi Univ. B.Sc. (Stat. Hons.), 1987*]

(c) Find by the method of likelihood ratio testing, a test for the null hypothesis $H_0 : m = m_0$ for a normal (m, σ^2) population, σ^2 known.

[*Calcutta Univ. B.Sc., (Maths Hons.), 1989*]

6. Discuss the general method of construction of likelihood ratio test.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\mu, \theta)$ population where θ is the unknown variance and μ is known. Obtain a likelihood ratio test for testing a simple $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1993*]

7. (a) Develop the likelihood ratio test for testing $H_0 : \mu = \mu_0$ based on a random sample of size n from $N(\mu, \sigma^2)$ population.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1982*]

(b) Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal population with mean μ and variance σ^2, μ and σ^2 being unknown. We wish to test $H_0 : \mu = \mu_0$ (specified) against $H_1 : \mu \neq \mu_0, 0 < \sigma^2 < \infty$.

Show that the Likelihood Ratio Test is same as the two tailed t-test.

[*Delhi Univ. M.A. (Eco.), 1986*]

(c) Describe the likelihood ratio test.

The random variable X follows normal distribution with mean θ_1 and variance θ_2. The parameter space is

$$\Theta = \{\theta_1, \theta_2\} : -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}.$$

Let $$\Theta_0 = \{(\theta_1, \theta_2) : \theta_1 = 0, 0 < \theta_2 < \infty\}.$$

Test the hypothesis $H_0 : \theta_1 = 0, \theta_2 > 0$ against the alternative composite hypothesis $H_1 : \theta_1 \neq 0, \theta_2 > 0$.

(*Madras Univ. B.Sc., 1988*)

8. Discuss the general method of construction of likelihood ratio test. Given $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, where all the parameters μ_1, μ_2, σ_1^2 and σ_2^2 are unspecified, develop the LR test for testing $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1983*]

9. Describe likelihood ratio test and state its important properties.

Let X_1 and X_2 be $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively where the means and variance are unspecified. Develop LR test for testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.

OR

Construct LR test for testing $H_0 : \theta = \theta_0$ against all its alternatives in $N(\theta, \sigma^2)$, where σ^2 is known.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

10. Show that the likelihood ratio test for testing the equality of variances of two normal distributions is the usual F-test.

11. Show that the likelihood ratio test for testing $H_0 : \alpha = 0$ against $H_1 : \alpha \neq 0$, based on a random sample of size n from

$$f(x ; \alpha ; \beta) = \frac{1}{2\beta} ; \alpha - \beta \leq x \leq \alpha + \beta$$

is $(R/2Z)^n$ where $R = X_{(n)} - X_{(1)}$ and $Z = \max\left[-X_{(1)}, X_{(n)}\right]$.

[*Delhi B.Sc. (Stat. Hons.), 1989, 1985*]

12. Show that the likelihood ratio principle leads to the same test, when testing a simple hypothesis against an alternative simple hypothesis, as that given by Neyman-Pearson theorem. [*Madras Univ. B.Sc., 1988*]

16·8. Non-parametric Methods. Most of the statistical tests that we have discussed so far had the following two features in common.

(*i*) The form of the frequency function of the parent population from which the samples have been drawn is assumed to be known, and

(*ii*) They were concerned with testing statistical hypothesis about the parameters of this frequency function or estimating its parameters.

For example, almost all the exact (small) sample tests of significance are based on the fundamental assumption that the parent population is normal and are concerned with testing or estimating the means and variances of these populations. Such tests, which deal with the parameters of the population are known as *Parametric Tests*. Thus, a parametric statistical test is a test whose model specifies certain conditions about the parameters of the population from which the samples are drawn.

On the other hand, a *Non-parametric (N.P.) Test* is a test that does not depend on the particular form of the basic frequency function from which the samples are drawn. In other words, non-parametric test does not make any assumption regarding the form of the population.

However, certain assumptions associated with N.P. tests are :

(*i*) Sample observations are independent.

(*ii*) The variable under study is continuous.

(*iii*) p.d.f. is continuous.

(*iv*) Lower order moments exist.

Obviously these assumptions are fewer and much weaker than those associated with parametric tests.

16·8·1. Advantages and Disadvantages of N.P. Methods over Parametric Methods. Below we shall give briefly the comparative study of parametric and non-parametric methods and their relative merits and demerits.

Advantages of N.P. Methods :

(*i*) N.P. methods are readily comprehensible, very simple and easy to apply and do not require complicated sample theory.

(*ii*) No assumption is made about the form of the frequency function of the parent population from which sampling is done.

(*iii*) No parametric technique will apply to the data which are mere classification (*i.e.*, which are measured in nominal scale), while N.P. methods exist to deal with such data.

(*iv*) Since the socio-economic data are not, in general, normally distributed, N.P. tests have found applications in Psychometry, Sociology and Educational Statistics.

(*v*) N.P. tests are available to deal with the data which are given in ranks or whose seemingly numerical scores have the strength of ranks. For instance, no parametric test can be applied if the scores are given in grades such as A^+, A^-, B, A, B^+, etc.

Disadvantages of N.P. Tests.

(*i*) N.P. tests can be used only if the measurements are nominal or ordinal. Even in that case, if a parametric test exists it is more powerful than the N.P. test. In other words, if all the assumptions of a statistical model are satisfied by the data and if the measurements are of required strength, then the N.P. tests are wasteful of time and data.

(*ii*) So far, no N.P. methods exist for testing interactions in 'Analysis of Variance' model unless special assumptions about the additivity of the model are made.

(*iii*) N.P. tests are designed to test statistical hypothesis only and not for estimating the parameters.

Remarks 1. Since no assumption is made about the parent distribution, the N.P. methods are sometimes referred to as *Distribution Free* methods. These tests are based on the 'Order Statistic' theory. In these tests we shall be using median, range, quartile, inter-quartile range, etc., for which an ordered sample is desirable. By saying that $x_1, x_2, \ldots, x_n$ is an ordered sample we mean $x_1 \leq x_2 \leq \ldots \leq x_n$.

2. The whole structure of the N.P. methods rests on a simple but fundamental property of order statistic, *viz.*

"The distribution of the area under the density function between any two ordered observations is independent of the form of the density function", which we shall now prove.

16·8·2. Basic Distribution. Let Z be a continuous random variable with a p.d.f. $f(.)$. Let $Z_1, Z_2, \ldots, Z_n$ be a random sample of size n from $f(.)$ and let $x_1, x_2, \ldots, x_n$ be the corresponding ordered sample. Then the joint density of $x_1, x_2, \ldots, x_n$ is given by

$$g(x_1, x_2, \ldots, x_n) = n\,!\, f(x_1) f(x_2) \ldots f(x_n), \quad -\infty < x_1 < x_2 < \ldots < x_n < \infty \qquad \ldots(16\cdot95)$$

the factor n ! appearing since there are n ! permutations of the sample observations and each gives rise to the same ordered sample.

Let us define

$$U_i = \int_{-\infty}^{x_i} f(z)\, dz = F(x_i), \; (i = 1, 2, \ldots, n) \qquad \ldots(16\cdot96)$$

where $F(.)$ is the distribution function of Z. But since $F(x_i)$ is a uniform random variable on [0, 1], U_i, ($i = 1, 2, \ldots, n$), defined in (16·96) are random variables following uniform distribution on [0, 1]. Thus the joint density $k(.)$ of the random variables U_i, ($i = 1, 2, \ldots, n$) is given by

$$k(u_1, u_2, \ldots, u_n) = n\,!, \; 0 \leq u_1 < u_2 < \ldots < u_k \leq 1 \qquad \ldots(16\cdot97)$$

and does not depend on $f(.)$.

$$E(U_i) = \int_0^1 \cdots \int_0^{u_3} \int_0^{u_2} u_i \, n! \; du_1 \, du_2 \ldots du_n$$

$$= \frac{i}{n+1} \qquad \text{(On simplification)} \qquad \ldots(16\cdot98)$$

Thus the expected area under $f(.)$ between two successive ordered observations is given by

$$E(U_i) - E(U_{i-1}) = \frac{i}{n+1} - \frac{i-1}{n+1} = \frac{1}{n+1}, \qquad \ldots(16\cdot98a)$$

which is independent of $f(.)$

16·8·3. Wald-Wolfowitz Run Test. Suppose $x_1, x_2, \ldots, x_{n_1}$ is an ordered sample from a population with density $f_1(.)$ and let $y_1, y_2, \ldots, y_{n_2}$ be an independent ordered sample from another population with density $f_2(.)$. We want to test if the samples have been drawn from the same population or from populations with the same density functions, *i.e.,* if $f_1(.) = f_2(.)$.

Let us combine the two samples and arrange the observations in order of magnitude to give the combined ordered sample as, (say),

$$x_1 \, x_2 \, y_1 \, y_2 \, y_3 \, x_3 \, y_4 \, x_4 \, x_5 \, y_5 \ldots \qquad \ldots(16\cdot99)$$

Run (*Definition*). *A run is defined as a sequence of letters of one kind surrounded by a sequence of letters of the other kind, and the number of elements in a run is usually referred to as the length (l) of the run.*

Thus in (16·99), we have in order, a run of x ($l = 2$), a run of y ($l = 3$), a run of x ($l = 1$), a run of y ($l = 1$) etc.

If both the samples come from the sample population then there would be thorough mingling of x's and y's and consequently the number of runs in the combined sample would be large. On the other hand if the samples come from two different populations so that their ranges do not overlap, then there would be only two runs of the type $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$. Generally, any difference in mean and variance would tend to reduce the number of runs. Thus the *alternative hypothesis will entail too few runs.*

Procedure. In order to test the Null Hypothesis $H_0 : f_1(.) = f_2(.)$ *i.e.,* the samples have come from the same population we count the number of runs 'U' in the combined ordered sample.

Null hypothesis is rejected if $U < u_0$, where the value of u_0 for given level of significance is determined from considering the distribution of U under H_0.

First of all let us find the probability of obtaining a specific arrangement (16·99) under $H_0 : f_1(.) = f_2(.) = f(.)$, (say).

If X's and Y's are transformed to U's and V's by the relation :

$$U_i = \int_{-\infty}^{x_i} f(z)\, dz, \qquad V_i = \int_{-\infty}^{y_i} f(z)\, dz,$$

then the joint *p.d.f.* of U's and V's becomes

$$g(u_1, u_2, \ldots, u_n, v_1, v_2, \ldots, v_n) = n_1 ! \, n_2 ! \qquad \ldots(16\cdot100)$$

The probability of an arrangement (16·99) is obtained on integrating (16·100) over the region defined by

$$0 < u_1 < u_2 < v_1 < v_2 < v_3 < \ldots < 1$$

i.e., integrating u_1 over 0 to u_2 ; then u_2 over 0 to v_1 and so on. The value of the integral will, on simplification, come out to be

$$\frac{n_1 ! \, n_2 !}{(n_1 + n_2) !} = \frac{1}{\binom{n_1 + n_2}{n_1}}$$

Since there are exactly $\binom{n_1 + n_2}{n_1}$ arrangements of n_1, x's and n_2, y's, it follows that all the arrangements of x's and y's are equally likely.

Since under H_0 all the $\binom{n_1 + n_2}{n_1}$ arrangements of n_1 x's and n_2 y's are equally likely, to obtain the distribution of U under H_0, it is necessary to count all the arrangements with exactly 'u' runs. Let us first take the case of even number of runs, *i.e.*, $u = 2k$. In this case we should have k runs of x's and k runs of y's.

n_1 x's will give k runs if they are separated by $(k - 1)$ vertical bars in distinct spaces between the x's. In other words, $(k - 1)$ spaces are to come out of the total number of $(n_1 - 1)$ spaces between the n_1 x's and this can happen in $\binom{n_1 - 1}{k - 1}$ ways. Hence k runs of x's can be obtained in $\binom{n_1 - 1}{k - 1}$ ways. Similarly, k runs of y's can be obtained in $\binom{n_2 - 1}{k - 1}$ ways.

The same result holds if the sequence of runs in (16·99) starts with x or with y. Since a sequence of type (16·99) may start with x or y, we get

$$P(U = 2k) = \frac{2\binom{n_1 - 1}{k - 1}\binom{n_2 - 1}{k - 1}}{\binom{n_1 + n_2}{n_1}}$$

If the number of runs in (16·99) is odd, *i.e.*, $u = 2k + 1$, then we should have either (*i*) $(k + 1)$ runs of x and k runs of y or (*ii*) k runs of x and $(k + 1)$ runs of y. Hence

$$\begin{aligned} P(U = 2k + 1) &= P(i) + P(ii) \\ &= \frac{\binom{n_1 - 1}{k}\binom{n_2 - 1}{k - 1} + \binom{n_1 - 1}{k - 1}\binom{n_2 - 1}{k}}{\binom{n_1 + n_2}{n_1}} \end{aligned}$$

Hence the distribution of U under H_0 is given by

$$\left.\begin{aligned} P(U=2k) &= \frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}, \text{ and} \\ P(U=2k+1) &= \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1}+\binom{n_1-1}{k-1}\binom{n_2-1}{k}}{\binom{n_1+n_2}{n_1}} \end{aligned}\right\} \qquad \ldots(16\cdot101)$$

If the probability of type I error is fixed as α, then u_0 is determined from the equation :

$$\sum_{u=2}^{u_0} h(u) = \alpha \qquad \ldots(16\cdot102)$$

where $h(u)$ is the probability function of U given by (16·101).

Calculation of u_0 from (16·102) is quite tedious and cumbersome unless n_1 and n_2 are large in which case under H_0, U is asymptotically normal with

$$E(U) = \frac{2n_1n_2}{n_1+n_2} + 1 \qquad \ldots(16\cdot103)$$

$$\text{Var}(U) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \qquad \ldots(16\cdot104)$$

and we can use the normal test

$$Z = \frac{U - E(U)}{\sqrt{\text{Var}(U)}} \sim N(0, 1), \text{ asymptotically.} \qquad \ldots(16\cdot105)$$

This approximation is fairly good if each of n_1 and n_2 is greater than 10. Since the alternative hypothesis is "too few runs', the test is ordinarily one-tailed with only negative values leading to the rejection of H_0.

16·8·4. Test for Randomness. Another application of the 'run' theory is in testing the randomness of a given set of observations. Let $x_1, x_2, \ldots, x_n$ be the set of observations arranged in the order in which they occur, *i.e.,* x_i is the *ith* observation in the outcome of an experiment. Then, for each of the observations, we see if it is above or below the value of the median of the observations and write A if the observation is above and B if it is below the median value. Thus we get a sequence of A's and B's of the type, (say),

$$A\ B\ B\ A\ A\ A\ B\ A\ B\ B \qquad \ldots(*)$$

Under the *null hypothesis H_0 that the set of observations is random,* the number of runs U in (*) is a r.v. with

$$E(U) = \frac{n+2}{2} \quad \text{and} \quad \text{Var}(U) = \frac{n}{4}\left(\frac{n-2}{n-1}\right) \qquad \ldots(16\cdot106)$$

For large n (say, > 25), U may be regarded as asymptotically normal and we may use the normal test.

16·8·5. Median Test. Median test is a statistical procedure for testing if two independent ordered samples differ in their central tendencies. In other words, it gives information if two independent samples are likely to have been drawn from the populations with the same median.

As in 'run' test, let $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ be two independent ordered samples from the populations with p.d.f.'s $f_1(.)$ and $f_2(.)$ respectively. The measurements must be at least ordinal. Let $z_1, z_2, \ldots, z_{n_1+n_2}$ be the combined *ordered* sample. Let m_1 be the number of x's and m_2 the number of y's exceeding the median value M, (say), of the combined sample.

Then under the null hypothesis that the samples come from the same population or from different populations with the same median, *i.e.*, under $H_0 : f_1(.) = f_2(.)$, the joint distribution of m_1 and m_2 is the hypergeometric distribution with probability function :

$$p(m_1, m_2) = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}}{\binom{n_1+n_2}{m_1+m_2}} \qquad \ldots(16\cdot107)$$

If $m_1 < n_1/2$, then the critical region corresponding to the size of type 1 error α, is given by $m_1 < m_1'$ where m_1' is computed from the equation

$$\sum_{m_1=1}^{m_1'} p(m_1, m_2) = \alpha \qquad \ldots(16\cdot108)$$

The distribution of m_1 under H_0 is also hyper-geometric with

$$\left.\begin{aligned} E(m_1) &= \frac{n_1}{2}, \text{ if } N = n_1 + n_2 \text{ is even} \\ &= \frac{n_1}{2} \cdot \frac{N-1}{N}, \text{ if } N \text{ is odd} \\ \text{and} \quad \text{Var}(m_1) &= \frac{n_1 n_2}{4(N-1)}, \text{ if } N \text{ is even} \\ &= \frac{n_1 n_2 (N+1)}{4N^2}, \text{ if } N \text{ is odd} \end{aligned}\right\} \qquad \ldots(16\cdot109)$$

This distribution is most of the times quite inconvenient to use. However for large samples, we may regard m_1 to be asymptotically normal and use normal test, *viz.*,

$$Z = \frac{m_1 - E(m_1)}{\sqrt{\text{Var}(m_1)}} \sim N(0, 1), \text{ asymptotically.} \qquad \ldots(16\cdot110)$$

Remarks 1. The observations m_1 and m_2 can be classified into the following 2×2 contingency table.

	Sample 1	Sample 2	Total
No. of observations $> M$	m_1	m_2	$m_1 + m_2$
No. of observations $< M$	$n_1 - m_1$	$n_2 - m_2$	$n_1 + n_2 - m_1 - m_2$
Total	n_1	n_2	$n_1 + n_2 = N$

If frequencies are small we can compute the exact probabilities from (16·107), rather than approximate them. However, if frequencies are large we may use χ^2-test with 1 d.f. (for a 2×2 contingency table) for testing H_0.

The approximation is fairly good if both n_1 and n_2 exceed 10.

2. Median test is sensitive to the differences in location between $f_1(x)$ and $f_2(y)$ but not to differences in their shapes. Thus if $f_1(x)$ and $f_2(y)$ have the same median, we would expect $H_0 : f_1(.) = f_2(.)$ to be accepted ordinarily even though their shapes are quite different.

3. Generally, the median test makes the correct decision with a little more assurance than does the sign test (*c.f.* §. 16·8·6) but not as decisively as the *t*-test.

16·8·6. Sign Test. Consider a situation where it is desired to compare two things or materials under various sets of conditions. An experiment is thus conducted under the following circumstances :

(*i*) When there are pairs of observations on two things being compared.

(*ii*) For any given pair, each of the two observations is made under similar extraneous conditions.

(*iii*) Different pairs are observed under different conditions.

Condition (*iii*) implies that the differences $d_i = x_i - y_i$; $i = 1, 2, \ldots, n$ have different variances and thus renders the paired *t*-test (Chapter 14) invalid, which would have otherwise been used unless there was obvious non-normality. So, in such a case we use the '***Sign Test***', named so since it is based on the signs (plus or minus) of the deviations $d_i = x_i - y_i$. No assumptions are made regarding the parent population. The only assumptions are :

(*i*) Measurements are such that the deviations $d_i = x_i - y_i$, can be expressed in terms of positive or negative signs.

(*ii*) Variables have continuous distribution.

(*iii*) d_i's are independent.

Different pairs (x_i, y_i) may be from different populations (say w.r.t. age, weight, stature, education, etc.). The only requirement is that within each pair, there is matching w.r.t relevant extraneous factors.

Procedure. Let (x_i, y_i), $i = 1, 2, \ldots, n$ be n paired sample observations drawn from the two populations with *p.d.f.*'s $f_1(.)$ and $f_2(.)$. We want to test the null hypothesis $H_0 : f_1(.) = f_2(.)$. To test H_0, consider $d_i = x_i - y_i$, $(i = 1, 2, \ldots, n)$. When H_0 is true, x_i and y_i constitute a random sample of size 2 from the same population. Since the probability that the first of the two sample observations exceeds the second is same as the probability that the second exceeds the first and since hypothetically the probability of a tie is zero, H_0 may be restated as :

$$H_0: \qquad P[X - Y > 0] = \tfrac{1}{2} \text{ and } P[X - Y < 0] = \tfrac{1}{2}$$

Let us define

$$U_i = \begin{cases} 1, & \text{if } x_i - y_i > 0 \\ 0, & \text{if } x_i - y_i < 0 \end{cases}$$

U_i is a Bernoulli variate with $p = P(x_i - y_i > 0) = \frac{1}{2}$. Since U_i's, $i = 1, 2, \ldots, n$ are independent, $U = \sum_{i=1}^{n} U_i$, the total number of positive deviations, is a Binomial variate with parameters n and p $(= \frac{1}{2})$. Let the number of positive deviations be k. Then

$$P(U \le k) = \sum_{r=0}^{k} \binom{n}{r} p^r q^{n-r}, \quad (p = q = \tfrac{1}{2} \text{ under } H_0).$$

$$= \left(\frac{1}{2}\right)^n \sum_{r=0}^{k} \binom{n}{r} = p', \text{ (say).} \qquad \ldots(16\cdot111)$$

If $p' \le 0{\cdot}05$, we reject H_0 at 5% level of significance and if $p' > 0{\cdot}05$, we conclude that the data do not provide any evidence against the null hypothesis, which may therefore, be accepted.

For large samples, $(n \ge 30)$, we may regard U to be asymptotically normal with, (under H_0)

$$E(U) = np = n/2 \quad \text{and} \quad \text{Var}(U) = npq = n/4.$$

$$\therefore \quad Z = \frac{U - E(U)}{\sqrt{\text{Var}(U)}} = \frac{U - n/2}{\sqrt{(n/4)}}, \text{ is asymptotically } N(0, 1), \qquad \ldots(16\cdot112)$$

and we may use normal test.

16·8·7. Mann-Whitney-Wilcoxon U-test. This non-parametric test for two samples was described by Wilcoxon and studied by Mann and Whitney. It is the most widely used test as an alternative to the t-test when we do not make the t-test assumptions about the parent population.

Let x_i $(i = 1, 2, \ldots, n_1)$ and y_j $(j = 1, 2, \ldots, n_2)$ be independent ordered samples of size n_1 and n_2 from the populations with *p.d.f.* $f_1(.)$ and $f_2(.)$ respectively. We want to test the null hypothesis $H_1 : f_1(.) = f_2(.)$. Like the run test, Mann-Whitney test is based on the pattern of the x's and y's in the combined ordered sample. Let T *denote the sum of ranks of the* y*'s in the* *combined ordered sample*. For example, for the pattern (16·99) on page 16·61 of combined ordered sample the ranks of y observations are respectively 3, 4, 5, 7, 10 etc. and $T = 3 + 4 + 5 + 7 + 10 + \ldots$ The test statistic U is then defined in terms of T as follows :

$$U = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - T \qquad \ldots(16\cdot113)$$

If T is significantly large or small then $H_0 : f_1(.) = f_2(.)$ is rejected. The problem is to find the distribution of T under H_0. Unfortunately, it is very troublesome to obtain the distribution of T under H_0. However, Mann and Whitney have obtained the distribution of T for small n_1 and n_2, have found the moments of T in general and shown that T is asymptotically normal. It has been established that under H_0, U is asymptotically normally distributed as $N(\mu, \sigma^2)$, where

$$\mu = E(U) = \frac{n_1 n_2}{2}$$

$$\sigma^2 = \text{Var}(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \qquad \ldots(16\cdot114)$$

Hence

$$Z = \frac{U - \mu}{\sigma} \sim N(0, 1), \text{ asymptotically}, \qquad \ldots(16\cdot114a)$$

and normal test can be used. The approximation is fairly good if both n_1 and n_2 are greater than 8.

Remark. The asymptotic relative efficiency (ARE) of Mann-Whitney's U-test relative to two samples t-test is greater than or equal to 0·864. For a normal population, this $ARE = 3/\pi = 0{\cdot}955$. Accordingly, Mann-Whitney's U-test is regarded as the best non-parametric test for location.

EXERCISE 16 (c)

1. Explain what is meant by non-parametric methods. How do they differ from parametric methods ? Illustrate your answer by considering a suitable nonparametric test for the hypothesis that two independent samples have come from the same population.

2. (*a*) Derive the sign test, stating clearly the assumptions made.

(*b*) Describe the median test for the two-sample location problem. Find the distribution of the test statistic and compute its mean and variance under the null hypothesis. How is the test carried out in case of large samples ?

3. Explain the main difference between parametric and non-parametric approaches to the theory of statistical inference. Derive the sign test for two sample problem. [*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

4. Describe the sign test.

$X_1, X_2, \ldots, X_{10}$ is a random sample of size 10 from a population having distribution function $F(x)$. Test the hypothesis $H_0 : F(72) = \frac{1}{2}$ against the alternative hypothesis, $H_1 : F(72) > \frac{1}{2}$. [*Madras Univ. B.Sc., 1988*]

5. Explain Median Test and how it is applied.

The observations of a random sample of size 10 from a distribution which is symmetric about $K_{.5}$ are 20·2, 24·1, 21·3, 17·2, 19·8, 16·5, 21·8, 18·7, 17·1, 19·9. Use Wilcoxon's Test to test the hypothesis $H_0 : K_{.5} = 18$ against $H_1 : K_{.5} > 18$ if $\alpha = 0{\cdot}05$. You may use the normal approximation.

[*Agra Univ. B.Sc., 1989*]

6. Describe the procedure in median test when there are two independent samples. What non-parametric test would you use when the two samples are related.

7. Discuss the Mann-Whitney-Wilcoxon test for the equality of two population distribution functions. [*Delhi Univ. B.Sc. (Stat. Hons.), 1986*]

8. What are the advantages and disadvantages of non-parametric methods over parametric methods ?

Develop the following non-parametric tests, stating the underlying assumptions and the null hypotheses :

(*a*) Median test

(*b*) Mann-Whitney-Wilcoxon test.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1993*]

9. Explain the main difference between the parametric and non-parametric approaches to the theory of statistical inference. What are the advantages of non-parametric tests ? Develop Median test and Mann-Whitney-Wilcoxon test.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1985*]

10. Distinguish between 'sign test' and 'Wilcoxon signed rank test'. Describe the sign test for testing that the population median is M_0 against the alternative that the median is M_1 ($> M_0$).

11. Develop the Mann-Whitney-Wilcoxon test and obtain the mean and variance of the test statistic T. How is the test carried out for large samples ?

12. Explaining the distinction between the parametric and non-parametric tests, write down the advantages of non-parametric tests. Also write their disadvantages.

Thirty observations as given below are obtained :

24, 35, 12, 50, 60, 70, 68, 49, 80, 25, 69, 28, 28, 11, 83,
31, 37, 34, 54, 75, 45, 95, 75, 26, 43, 57, 94, 48, 63, 45

Test their randomness by considering the sequence of positive and negative signs. [*Agra Univ. B.Sc., 1989*]

13. What are the advantages and disadvantages of Non-Parametric Methods over Parametric Methods ?

Derive the Wald-Wolfowitz run test for testing the equality of two distribution functions. [*Delhi Univ. B.Sc. (Stat. Hons.), 1987*]

14. What are the advantages of Non Parametric tests ? Define a run and the length of a run. Describe the Run Test in detail for testing the equality of the two populations and extend the test when the ties occur.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1983*]

15. What are runs ? Comment on their utility in non-parametric inference.

If R_1 and R_2 denote the number of runs of n_1 objects of one type and n_2 objects of another type in a sample of size $n_1 + n_2$, then find the probability that $R_1 + R_2 = r$, for r even and r odd, and also the mean and variance of $R_1 + R_2$ when all these $n_1 + n_2$ observations arise from the same distribution.

[*Delhi Univ. M.Sc. (Stat.), 1991*]

16. (*i*) Explain how the run test can be used to test randomness.

(*ii*) In the median test with samples of size 9 and 7 respectively, from two populations find the probability density function of the random variable representing the number of values of the samples from the first population in the lower half of the combined sample. [*Madras Univ. B.Sc., 1988*]

17. (*a*) The win-lose record of a certain basketball team for its last 50 consecutive games was as follows :—

W W W W W W L W W W W W W L W L W W W L L W W W W
L W W W L L W W W W W W L L W W L L L W W L W W W

Apply run test to test that sequence of wins and losses is random.

(*b*) Use an appropriate non-parametric test procedure to test for randomness the following set of 30 two-digit numbers :

15, 17, 01, 65, 69, 69, 58 41, 81, 16,
16, 20, 00, 84, 22, 28, 26, 46, 66, 36,
86, 66, 17, 34, 49, 85, 45, 51, 40, 10.

18. At the beginning of the year a first grade class was randomly divided into two groups. One group was taught to read using a uniform method, where all students progressed from one stage to the next at the same time, following the teacher's direction. The second group was taught to read using an individual method, where each student progressed at his own rate according to a programmed work book, under supervision of the teacher. At the end of the year each student was given a reading ability test with the following results :

First Group			*Second Group*		
227	55	184	202	271	63
176	234	147	14	151	284
252	194	88	165	235	53
149	247	161	171	147	228
16	92	171	292	99	271

Use the Wald-Wolfowitz run-test to test for the equality of the distribution functions of the two groups.

19. Using the number of runs above and below the median, test for randomness the following set of a table of 2-digit numbers :

15, 77, 01, 65, 69, 69, 58, 40, 81, 16, 16 20, 00, 84, 22,
28, 26, 46, 66, 36, 86, 66, 17, 43, 49, 85, 40, 51, 40, 10.

16·9. Sequential Analysis – Introduction We have seen that in Neyman-Pearson theory of testing hypothesis, n, the sample size is regarded as a fixed constant and keeping α fixed, we minimise β. But in the sequential analysis theory propounded by *A. Wald* n, the sample size is not fixed but is regarded as a random variable whereas both α and β are fixed constants.

16·9·1. Sequential Probability Ratio Test (SPRT). The best known procedure in sequential testing is the *Sequential Probability Ratio Test* (SPRT) developed by A. Wald discussed below.

Suppose we want to test the hypothesis, $H_0 : \theta = \theta_0$ against the alternative hypothesis, $H_1 : \theta = \theta_1$, for a distribution with p.d.f. $f(x, \theta)$ For any positive integer m, the likelihood function of a sample $x_1, x_2, \ldots, x_m$ from the population with p.d.f. $f(x, \theta)$ is given by

$$L_{1m} = \prod_{i=1}^{m} f(x_i, \theta_1) \text{ when } H_1 \text{ is true,}$$

and by $$L_{0m} = \prod_{i=1}^{m} f(x_i, \theta_0) \text{ when } H_0 \text{ is true,}$$

and the likelihood ratio λ_m is given by

$$\lambda_m = \frac{L_{1m}}{L_{0m}} = \frac{\prod_{i=1}^{m} f(x_i, \theta_1)}{\prod_{i=1}^{m} f(x_i, \theta_0)} = \prod_{i=1}^{m} \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)}, \ (m = 1, 2, \ldots) \qquad \ldots(16\cdot115)$$

The SPRT for testing H_0 against H_1 is defined as follows :

At each stage of the experiment (at the mth trial for any integral value m), the likelihood ratio λ_m, $(m = 1, 2, \ldots)$ is computed.

$$\left.\begin{array}{ll} (i) & \text{If } \lambda_m \geq A, \text{ we terminate the process with the rejection of } H_0 \\ (ii) & \text{If } \lambda_m \leq B, \text{ we terminate the process with the acceptance of } H_0, \text{ and} \\ (iii) & \text{If } B < \lambda_m < A, \text{ we continue sampling by taking an additional observation.} \end{array}\right\} \qquad \ldots(16\cdot116)$$

Here A and B, $(B < A)$ are the constants which are determined by the relation

$$A = \frac{1-\beta}{\alpha}, \ B = \frac{\beta}{1-\alpha} \qquad \ldots(16\cdot117)$$

where α and β are the probabilities of type I error and type II error respectively.

From computational point of view, it is much convenient to deal with $\log \lambda_m$ rather than λ_m, since

$$\log \lambda_m = \sum_{i=1}^{m} \log \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)} = \sum_{i=1}^{m} z_i \qquad \ldots(16\cdot118)$$

where $$z_i = \log \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)} \qquad \ldots(16\cdot118a)$$

In terms of z_i's, $SPRT$ is defined as follows :

$$\left.\begin{array}{ll} (i) & \text{If } \sum z_i \geq \log A, \text{ reject } H_0 \\ (ii) & \text{If } \sum z_i \leq \log B, \text{ reject } H_1 \\ (iii) & \text{If } \log B < \sum z_i < \log A, \text{ continue sampling by taking an additional observation.} \end{array}\right\} \qquad \ldots(16\cdot119)$$

Remarks 1. In *SPRT*, we continue taking additional observations unless the inequality

$$B < \lambda_m < A \quad \Rightarrow \quad \log B < \sum z_i < \log A,$$

is violated at either end. It has been proved that SPRT eventually terminates with probability one.

2. Sequential schemes provide for a minimum amount of sampling and thus result is considerable saving in terms of inspection, time and money. As compared with single sampling, sequential scheme requires on the average 33% to 50% less inspection for the same degree of protection *i.e.,* for the same values of α and β.

16·9·2. Operating Characteristic (O.C.) Function of SPRT. The O.C. function $L(\theta)$ is defined as

$L(\theta) =$ Probability of accepting $H_0 : \theta = \theta_0$ when θ is the true value of the parameter,

and since the power function

$P(\theta) =$ Probability of rejecting H_0 where θ is the true value, we get

$$L(\theta) = 1 - P(\theta) \qquad \ldots(16·120)$$

The O.C. function of a SPRT for testing $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$, in sampling from a population with density function $f(x, \theta)$ is given by

$$L(\theta) = \frac{A^{h(\theta)} - 1}{A^{h(\theta)} - B^{h(\theta)}}, \qquad \ldots(16·121)$$

where for each value of θ, the value of $h(\theta) \neq 0$, is to be determined so that

$$E\left[\frac{f(x, \theta_1)}{f(x, \theta_0)}\right]^{h(\theta)} = 1 \qquad \ldots(16·122)$$

where the constants A and B have already been defined in (16·117). It has been proved that under very simple conditions on the nature of the function $f(x, \theta)$, there exists a unique value of $h(\theta) \neq 0$ such that (16·122) is satisfied.

16·9·3. Average Sample Number (A.S.N.). The sample size n in sequential testing is a random variable which can be determined in terms of the true density function $f(x, \theta)$. The A.S.N. function for the S.P.R.T. for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, is given by

$$E(n) = \frac{L(\theta) \log B + [1 - L(\theta)] \log A}{E(Z)} \qquad \ldots(16·123)$$

where $$Z = \log\left[\frac{f(x, \theta_1)}{f(x, \theta_0)}\right], A = \frac{1 - \beta}{\alpha}, B = \frac{\beta}{1 - \alpha} \qquad \ldots(16·123a)$$

Example 16·11. *Give the S.P.R.T. for testing* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta = \theta_1 \ (> \theta_0)$, *in sampling from a normal density.*

$$\frac{x}{\theta} = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x - \theta}{\sigma}\right)^2\right], -\infty < x < \infty$$

where σ *is known. Also obtain its O.C. function and A.S.N. function.*

Solution. $\dfrac{f(x_i, \theta_1)}{f(x_i, \theta_0)} = \exp\left[-\dfrac{1}{2\sigma^2}\{(x_i-\theta_1)^2-(x_i-\theta_0)^2\}\right]$

$$= \exp\left[-\frac{1}{2\sigma^2}\{(\theta_0-\theta_1)(2x_i-\theta_0-\theta_1)\}\right] \qquad \ldots(*)$$

$$\therefore \qquad z_i = \log\frac{f(x_i, \theta_1)}{f(x_i, \theta_0)} = \frac{\theta_1-\theta_0}{\sigma^2}\left[x_i - \frac{\theta_0+\theta_1}{2}\right] \qquad \ldots(**)$$

$$\Rightarrow \qquad \log\lambda_m = \sum_{i=1}^{m} z_i = \frac{\theta_1-\theta_0}{\sigma^2}\left[\sum_i x_i - \frac{m(\theta_0+\theta_1)}{2}\right]$$

Hence the S.P.R.T. for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, is given by [c.f. (16·119)] :

(*i*) Reject H_0 if

$$\frac{\theta_1-\theta_0}{\sigma^2}\left[\Sigma x_i - \frac{m(\theta_0+\theta_1)}{2}\right] \geq \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\Rightarrow \qquad \sum_{i=1}^{m} x_i \geq \frac{\sigma^2}{\theta_1-\theta_0}\log\left(\frac{1-\beta}{\alpha}\right) + \frac{m(\theta_0+\theta_1)}{2}; \ (\theta_1 > \theta_0)$$

(*ii*) Accept H_0 if

$$\frac{\theta_1-\theta_0}{\sigma^2}\left[\Sigma x_i - \frac{m(\theta_0+\theta_1)}{2}\right] \leq \log\left(\frac{\beta}{1-\alpha}\right)$$

$$\Rightarrow \qquad \sum_{i=1}^{m} x_i \leq \frac{\sigma^2}{\theta_1-\theta_0}\log\left(\frac{\beta}{1-\alpha}\right) + \frac{m(\theta_0+\theta_1)}{2}; \ (\theta_1 > \theta_0)$$

and (*iii*) Continue taking additional observations as long as

$$\log\left(\frac{\beta}{1-\alpha}\right) < \frac{\theta_1-\theta_0}{\sigma^2}\left[\Sigma x_i - \frac{m(\theta_0+\theta_1)}{2}\right] < \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\Rightarrow \frac{\sigma^2}{\theta_1-\theta_0}\log\left(\frac{1-\beta}{\alpha}\right) + \frac{m(\theta_0+\theta_1)}{2} < \Sigma x_i < \frac{\sigma^2}{\theta_1-\theta_0}\log\left(\frac{\beta}{1-\alpha}\right) + \frac{m(\theta_0+\theta_1)}{2}$$

O.C. Function. First of all we shall determine $h = h(\theta) \neq 0$, from (16·122) *i.e.*, from

$$\int_{-\infty}^{\infty}\left[\frac{f(x, \theta_1)}{f(x, \theta_0)}\right]^h f(x, \theta)\, dx = 1$$

$$\Rightarrow \quad \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2\right] \cdot \left[\exp\left\{-\frac{1}{2\sigma^2}(\theta_1-\theta_0) \times (-2x+\theta_0+\theta_1)\right\}\right]^h dx = 1, \text{ [On using (*)]}$$

$$\Rightarrow \quad \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left\{x^2 - 2x\left((\theta_1-\theta_0)h+\theta\right) + \theta^2 + (\theta_1^2-\theta_0^2)h\right\}\right] dx = 1$$

If we take

$$\left.\begin{aligned} \lambda &= (\theta_1-\theta_0)\,h+\theta \\ \lambda^2 &= (\theta_1^2-\theta_0^2)h+\theta^2 \end{aligned}\right\} \qquad \ldots(***)$$

then L.H.S. becomes

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x-\lambda)^2\right]dx,$$

which being the total area under normal probability curve with mean λ and variance σ^2 is always unity, as desired. Thus $h = h(\theta)$ is the solution of (***) and is given by

$$(\theta_1^2-\theta_0^2)\,h+\theta^2 = [\theta_1-\theta_0)h+\theta]^2$$

$$\Rightarrow \quad (\theta_1^2-\theta_0^2)h = (\theta_1-\theta_0)^2h^2+2\theta(\theta_1-\theta_0)h$$

Since $h = h(\theta) \neq 0$ and $\theta_1 \neq \theta_0$, on dividing throughout by $(\theta_1-\theta_0)h$, we get

$$(\theta_1+\theta_0) = (\theta_1-\theta_0)\,h+2\theta$$

$$\Rightarrow \quad h(\theta) = \frac{\theta_1+\theta_0-2\theta}{\theta_1-\theta_0}$$

Substituting for $h(\theta)$ in (16·121) we get the required expression for the O.C. function.

A.S.N. function. We have

$$Z = \log\frac{f(x,\theta_1)}{f(x,\theta_0)} = \frac{\theta_1-\theta_0}{\sigma^2}\left[x-\frac{\theta_0+\theta_1}{2}\right] \qquad \text{[From (**)]}$$

$$\therefore \quad E(Z) = \frac{\theta_1-\theta_0}{2\sigma^2}\left[2E(x)-\theta_0-\theta_1\right]$$

$$= \frac{\theta_1-\theta_0}{2\sigma^2}\left[2\theta-\theta_0-\theta_1\right]$$

Substituting in (16·123), we get the required A.S.N. function.

Example 16·12. *Let X have the distribution :*

$$f(x,\theta) = \theta^x(1-\theta)^{1-x}\,;\; x = 0, 1;\; 0 < \theta < 1$$

For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, construct S.P.R.T. and obtain its A.S.N. and O.C. functions.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1993, 1985*]

Solution. We have

$$\lambda_m = \frac{L(x_1, x_2, \ldots, x_m \mid H_1)}{L(x_1, x_2, \ldots, x_m \mid H_0)}$$

$$= \left\{ \theta_1^{\sum_{i=1}^{m} x_i} (1-\theta_1)^{m - \sum_{i=1}^{m} x_i} \right\} \div \left(\theta_0^{\sum x_i} (1-\theta_0)^{m - \sum x_i} \right)$$

$$= \left(\frac{\theta_1}{\theta_0} \right)^{\sum_{i=1}^{m} x_i} \left(\frac{1-\theta_1}{1-\theta_0} \right)^{m - \sum_{i=1}^{m} x_i}$$

$$\log \lambda_m = \sum x_i \log (\theta_1/\theta_0) + (m - \sum x_i) \log \left(\frac{1-\theta_1}{1-\theta_0} \right)$$

$$= \sum_{i=1}^{m} x_i \log \left[\frac{\theta_1 (1-\theta_0)}{\theta_0 (1-\theta_1)} \right] + m \log \left(\frac{1-\theta_1}{1-\theta_0} \right)$$

Hence $SPRT$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, is given by [*c.f.* (16·119)] :

(*i*) Accept H_0 if $\log \lambda_m \le \log \left(\dfrac{\beta}{1-\alpha} \right) = b$, (say)

i.e., if $\sum_{i=1}^{m} x_i \le \dfrac{b - m \log [(1-\theta_1)/(1-\theta_0)]}{\log [\theta_1 (1-\theta_0)/\theta_0 (1-\theta_1)]} = a_m$, (say).

(*ii*) Reject H_0 (Accept H_1) if $\log \lambda_m \ge \log \dfrac{1-\beta}{\alpha} = a$, (say)

i.e., if $\sum_{i=1}^{m} x_i \ge \dfrac{a - m \log [(1-\theta_1)/(1-\theta_0)]}{\log [\theta_1 (1-\theta_0)/\theta_0 (1-\theta_1)]} = r_m$, (say).

(*iii*) Continue sampling if

$$b < \log \lambda_m < a \quad \Rightarrow \quad a_m < \sum x_i < r_m$$

O.C. Function. O.C. function is given by :

$$L(\theta) = [A^{h(\theta)} - 1] / [A^{h(\theta)} - B^{h(\theta)}] \qquad [c.f.\ (16\cdot121)] \quad \ldots(i)$$

where for each value of θ, $h(\theta) \ne 0$ is to be determined such that

$$E \left[\frac{f(x, \theta_1)}{f(x, \theta_0)} \right]^{h(\theta)} = 1 \qquad [c.f.\ (16\cdot122)]$$

$$\Rightarrow \quad \sum_{x=0}^{1} \left[\frac{f(x, \theta_1)}{f(x, \theta_0)} \right]^{h(\theta)} f(x, \theta) = 1$$

$$\Rightarrow \quad \sum_{x=0}^{1} \left[\left(\frac{\theta_1}{\theta_0} \right)^x \left(\frac{1-\theta_1}{1-\theta_0} \right)^{1-x} \right]^{h(\theta)} \theta^x (1-\theta)^{1-x} = 1$$

$$\Rightarrow \quad \left(\frac{1-\theta_1}{1-\theta_0} \right)^{h(\theta)} . (1-\theta) + \left(\frac{\theta_1}{\theta_0} \right)^{h(\theta)} . \theta = 1 \qquad \ldots(ii)$$

The solution of this equation for $h = h(\theta)$ is very tedious. From practical point of view, instead of solving (*ii*) for h we regard h as a parameter and solve it for θ, thus giving

$$\theta\left[\left(\frac{\theta_1}{\theta_0}\right)^{h(\theta)} - \left(\frac{1-\theta_1}{1-\theta_0}\right)^{h(\theta)}\right] = 1 - \left(\frac{1-\theta_1}{1-\theta_0}\right)^{h(\theta)}$$

$$\Rightarrow \qquad \theta = \frac{1 - [(1-\theta_1)/(1-\theta_0)]^{h(\theta)}}{(\theta_1/\theta_0)^{h(\theta)} - [(1-\theta_1)/(1-\theta_0)]^{h(\theta)}} = \theta(h), \text{ (say)}. \quad \ldots(iii)$$

Using (*i*), we have

$$L(\theta) = \frac{[(1-\beta)/\alpha]^h - 1}{[(1-\beta)/\alpha]^h - [\beta/(1-\alpha)]^h} = L(\theta, h), \text{ (say)}. \quad \ldots(iv)$$

Various points on the *O.C.* curve are obtained by assigning arbitrary values to 'h' and computing the corresponding values of θ and $L(\theta)$ from (*iii*) and (*iv*) respectively.

A.S.N. Function.

$$Z = \log\left[\frac{f(x,\theta_1)}{f(x,\theta_0)}\right];\ A = \frac{1-\beta}{\alpha},\ B = \frac{\beta}{1-\alpha}$$

$$\therefore \qquad E(Z) = \sum_{x=0}^{1} \log\left[\frac{f(x,\theta_1)}{f(x,\theta_0)}\right].f(x,\theta)$$

$$= \sum_{x=0}^{1} \log\left[\left(\frac{\theta_1}{\theta_0}\right)^x \left(\frac{1-\theta_1}{1-\theta_0}\right)^{1-x}\right].\theta^x(1-\theta)^{1-x}$$

$$= (1-\theta)\log\left(\frac{1-\theta_1}{1-\theta_0}\right) + \theta\,.\log\left(\frac{\theta_1}{\theta_0}\right)$$

$$= \theta\log\left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right] + \log\left(\frac{1-\theta_1}{1-\theta_0}\right) \quad \ldots(v)$$

A.S.N. is given by

$$E(n) = \frac{L(\theta)\log B + [1 - L(\theta)]\,.\log A}{E(Z)} \quad \ldots(vi)$$

Substituting the values of $E(Z)$ and $L(\theta)$ from (*v*) and (*iv*) in (*vi*), we get the *A.S.N.* function.

Remark. If h assumes negative values *i.e.*, if instead of h we take $-h$ where $h > 0$, then

$$L(\theta, -h) = \frac{A^{-h} - 1}{A^{-h} - B^{-h}} = \left(\frac{1 - A^h}{B^h - A^h}\right)B^h = \left(\frac{A^h - 1}{A^h - B^h}\right).B^h$$

$$\Rightarrow \qquad L(\theta, -h) = B^h.L(\theta, h) \quad \ldots(vii)$$

and

$$\theta(-h) = \frac{[(1-\theta_1)/(1-\theta_0)]^h - 1}{[(1-\theta_1)/(1-\theta_0)]^h - (\theta_1/\theta_0)^h}\left(\frac{\theta_1}{\theta_0}\right)^h$$

$$= \theta(h).\left(\frac{\theta_1}{\theta_0}\right)^h \quad \ldots(viii)$$

Formulae (*vii*) and (*viii*) are very convenient to use for obtaining the points on O.C. curve for arbitrary negative values of h.

EXERCISE 16(d)

1. (*a*) Describe Wald's Sequential Probability Ratio Test.

(*b*) Explain how the sequential test procedure differs from the Neyman-Pearson test procedure.

2. Define the OC function and ASN function in sequential analysis. Derive their approximate expressions for the sequential probability ratio test of a simple hypothesis against a simple alternative.

3. Describe Wald's S.P.R.T. Let X be a Bernoulli variate with p.d.f.

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}\,;\,x = 0, 1;\,0 \le \theta \le 1.$$

Employ S.P.R.T. for testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$, and obtain its A.S.N. and O.C. functions.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1993, '85*]

4. (*a*) Explain how the sequential test procedure differs from the Neyman-Pearson test procedure.

Develop the S.P.R.T. for testing $H_0: \pi = \pi_0$ against $H_1: \pi = \pi_1$, based on a random sample from a binomial population with parameters (n, π), n being known. Obtain its O.C. and A.S.N. functions.

(*b*) Obtain the sequential probability ratio test of the hypothesis $H_0: \theta = \frac{1}{3}$ against $H_1: \theta = \frac{2}{3}$ for the distribution :

$$f(x;\theta) = \begin{cases} \theta^x(1-\theta)^{1-x}, & \text{for } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

[*Madras Univ. B.Sc., 1988*]

5. Develop S.P.R. test for testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$, $(\theta_1 > \theta_0)$, where θ is the parameter of a Poisson distribution. Find approximate expressions for OC function and ASN function of the test.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

6. Describe S.P.R.T., its OC and ASN functions.

Construct S.P.R.T. for testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$, $(0 < \theta_0 < \theta_1)$, on the basis of a random sample drawn from the Pareto distribution with density function :

$$f(x, \theta) = \frac{\theta a^\theta}{x^{\theta+1}}, \quad x \ge a.$$

Also obtain its O.C. function and A.S.N. function.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

7. Explain how the sequential test procedure differs from the Neyman-Pearson test procedure.

Develop the S.P.R.T. for testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1\ (> \theta_0)$, based on a random sample of size n from a population with p.d.f.

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}, x > 0, \theta > 0.$$

Also obtain its $A.S.N.$ and $O.C.$ functions.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1987*]

8. Let X have the p.d.f.

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x} ; x \geq 0, \theta > 0 \\ 0 \quad , \text{elsewhere} \end{cases}$$

For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, construct the $S.P.R.T.$ and obtain its ASN and OC functions.

[*Indian Civil Services (Main), 1989;*
Delhi Univ. B.Sc. (Stat. Hons.), 1982, 1986]

9. (*a*) What is a sequential test ? How will you develop an optimum test of a specified strength for a simple null hypothesis versus a simple alternative ?

(*b*) Find expressions for the sample size expected for termination of $SPRT$ both under H_0 and H_1. Clearly state all the assumptions made.

(*c*) A random variable follows the normal distribution $N [\theta, \sigma^2]$, where σ^2 is known. Derive the $SPRT$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Obtain the approximate expression for the OC function.

[*Indian Civil Services (Main), 1990*]

10. To test sequentially the hypothesis H_0 that the distribution is given by $P(X = -1) = P(X = 1) = P(X = 2) = \frac{1}{3}$ against the alternative H_1 that it is given by $P(X = -1) = P(X = 1) = \frac{1}{4}$; $P(X = 2) = \frac{1}{2}$, it is decided to continue sampling as long as $-\frac{(n + 1)}{2} < S_n < \frac{n + 2}{2}$, where $S_n = X_1 + X_2 + \ldots + X_n$, the X_k's being the successive observations. Compute the probability under H_0 and under H_1 that the procedure will terminate with the fourth observation or earlier.

11. $X_1, X_2, \ldots, X_n$ be a sequence of $i.i.d.$ observations from $N(\mu, \sigma^2)$, where μ is known, σ^2 being unknown. Obtain the $SPRT$ for testing $H_0 : \sigma^2 = \sigma_0^2$, against $H_1 : \sigma^2 = \sigma_1^2$ ($> \sigma_0^2$). Also obtain its OC function and $A.S.N.$ function.

ADDITIONAL EXERCISE ON CHAPTER XVI

1. (*a*) "An examiner may pass a dull student or may fail a good student". Explain the above statement with reference to type-I and type-II errors.

2. A single value x is drawn from a normal population with mean m and variance 25. The null hypothesis $H_0 : m = 50$ is accepted if $x \leq 75$, otherwise $H_1 : m = 60$ is considered true. Evaluate the type I and type II errors.

3. Let p be the proportion of smokers in a certain city. You desire to test the hypothesis $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{3}{4}$. If you reject H_0 when 60 persons or more are found smokers in a sample of 100 persons, compute the significance level and power of the test.

4. Let $X_1, X_2, \ldots, X_{20}$ be a random sample of size 20 from a Poisson distribution with mean θ. Show that the critical region defined by $\sum_1^{20} x_i \geq 5$, is a uniformly most powerful critical region for testing $H_0 : \theta = 1/10$ against $H_1 : \theta > 1/10$.

5. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution $N(\theta, 16)$. Find the sample size n and a uniformly most powerful test of $H_0 : \theta = 25$ against $H_1 : \theta < 25$, with power function $K(\theta)$ so that approximately $K(25) = 0{\cdot}10$ and $K(23) = 0{\cdot}90$.

6. In testing $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma = \sigma_1$ ($\neq \sigma_0$), for the distribution :

$$f(x) = \frac{1}{\sigma} \exp\left[-\left(\frac{x-\theta}{\sigma}\right)\right], \ (\theta \leq x < \infty, \sigma > 0)$$

show that the UMP test is of the form

$$\Sigma x_i \geq \text{constant} \ \text{ and } \ \Sigma x_i \leq \text{constant}.$$

7. X_1, X_2 is a random sample from a distribution with p.d.f. $f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}$, $x > 0, \theta > 0$. The hypothesis $H_0 : \theta = 2$ is tested against $H_1 : \theta > 2$ and is rejected if and only if $X_1 + X_2 \geq 9{\cdot}5$. Obtain the power function and the significance level of the test. Also find the probability of type II error when $\theta = 4$.

8. On the basis of a single observation x from the following p.d.f.

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \ (x > 0; \theta > 0)$$

the null hypothesis, $H_0 : \theta = 1$ against the alternative hypothesis $H_1 : \theta = 4$, is tested by using a set

$$C = \{x : x > 3\}$$

as the critical region. Prove that the critical region C provides a most powerful test of its size. What is the power of the test ?

9. Let X be a single observation from the density $f(x\,;\theta) = 2\theta x + 1 - \theta$, $0 < x < 1$, $|\theta| \leq 1$; zero otherwise. Find the best critical region of size α, for testing $H_0 : \theta = 0$ against $H_1 : \theta < 0$. Express the power function of this test in terms of α. Is the test uniformly most powerful ? Explain.

10. $X_1, X_2, \ldots, X_n$ is a random sample of size n from $N(\theta, 100)$. For testing $H_0 : \theta = 75$ against $H_1 : \theta > 75$, the following test procedure is proposed :

$$\text{Reject } H_0 \text{ if } \bar{x} \geq c; \ \text{ Accept } H_0 \text{ if } \bar{x} < c.$$

Determine n and c so that the power function $P(\theta)$ of the test satisfies

$$P(75) = 0{\cdot}159 \text{ and } P(77) = 0{\cdot}841.$$

11. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution having p.d.f.

$$f(x, \theta) = \frac{[x(1-x)]^{\theta - 1}}{B\,(\theta, \theta)}, \ 0 < x < 1, \theta > 0$$
$$= 0\,, \ \text{elsewhere}$$

Show that the best critical region for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$ is

$$C = \left\{ (x_1, x_2, \ldots, x_n) : c \leq \prod_{i=1}^{n} x_i (1 - x_i) \right\}.$$

12. Let X be a single observation from the distribution with p.d.f.

$$f(x : \theta) = \theta . e^{-\theta x} ; \quad 0 < x < \infty, (\theta > 0)$$
$$= 0 , \quad \text{elsewhere.}$$

Obtain the best critical region of size α for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$. Also obtain the power of this test.

[Delhi Univ. M.Sc. (Maths), 1990]

13. Let $(x_1, x_2, \ldots, x_9)$ be a random sample from $N(\mu, 9)$. To test the hypothesis $H_0 : \mu = 40$ against $H_1 : \mu \neq 40$, consider the following two critical regions :

$$C_1 = \{ \bar{x} : \bar{x} \geq a_1 \}$$

$$C_2 = \{ \bar{x} : | \bar{x} - 40 | \geq a_2 \}$$

(*i*) Obtain the values of a_1 and a_2 so that the size of each critical region is 0·05.

(*ii*) Calculate the power of the two critical regions when $\mu = 39$ and $\mu = 41$ and comment on the results.

[Delhi Univ. M.A. (Eco.), 1992]

14. (*a*) For a sample of size 25 from a normal population $N(\mu, 25)$, $\bar{X} = 11·5$. Test the hypothesis $H_0 : \mu = 10$ against the alternative $H_1 : \mu > 10$. Calculate the power of the test for $\mu = 11$.

[Delhi Univ. M.A. (Eco.), 1988]

(*b*) Let $X \sim N(\mu, 25)$. The null and alterantive hypotheses are :

$$H_0 : \mu = 10 \text{ and } H_1 : \mu < 10$$

(*i*) Give the best test of size $\alpha = 0·05$ for a sample of size 25. (No derivation is expected).

(*ii*) Calculate the power of the test for $\mu = 8$.

[Delhi Univ. M.A. (Eco.), 1990]

15. X is normally distributed with $\sigma = 5$ and it is desired to test $H_0 : \mu = 105$ against $H_1 : \mu = 110$. How large a sample shoud be taken if the probability of accepting H_0 when H_1 is true is 0·02 and if a critical region of size 0·05 is used ?

(Agra Univ. B.Sc. 1989)

16. Let p be the probability that a given die shows an even number. To test $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{1}{3}$; the following procedure is adopted. Toss the die twice and accept H_0 if both times it shows even number. Find the probabilities of type I and type II errors.

(Delihi Univ. M.C.A., 1990)

17. The p.d.f. of x is given by $f(x) = \dfrac{1}{\theta}$, $0 < x < \theta$. Let the null hypothesis be $H_0 : \theta = \frac{4}{3}$ against the alternative hypothesis $H_A : \theta > \frac{4}{3}$. We have a random sample of one observation. The critical region is defined by $C = \{x : x > 1\}$.

(*i*) Find the significance level of the test.

(*ii*) Find the power of the test for $\theta = 7/3$ and $\theta = 10/3$.

[*Delhi Univ. M.A. (Eco.), 1991*]

TABLE I
LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	·0000	0043	0086	0120	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	·0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	·1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	·1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	·1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	·2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	·2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	·2553	2577	2601	2625	2648	2672	2695	2618	2742	2765	2	5	7	9	12	14	16	19	21
19	·2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	·3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	·3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	·3424	3444	3464	3483	3502	3522	3541	3562	3579	3598	2	4	6	8	10	12	14	15	17
23	·3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	·3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	·3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	·4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	·4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	·4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	·4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	·4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	·4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	·5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	·5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	·535	5315	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	·5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	·5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	·5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	·5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	·5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	·6021	6031	6042	6053	6065	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	·6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	·6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	·6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	·6435	6444	6454	6465	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	·6532	6542	6551	6561	6571	6380	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	·6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	·6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	·6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	·6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	·6990	6993	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	·7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	·7160	7168	7177	7185	7193	7202	7210	7218	7225	7235	1	2	2	3	4	5	6	7	7
53	·7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	·7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7

TABLE·I
LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
55	·7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	·74 2	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	·7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	·7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	·7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	·7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	·7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	·7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	·7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	·8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	·8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	·8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	·8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	·8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	·8388	8395	8401	8407	8414	8420	8426	8432	8439	3445	1	1	2	2	3	4	4	5	6
70	·8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	6
72	·8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	6
73	·8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1		2	2	3	4	5	6
74	·8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	6
75	·8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	6
76	·8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	6
77	·8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	·8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	·9331	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	·9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	·9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	·9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	[illegible]
84	·9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	·9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	·9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	·9395	9400	9405	9110	9415	9220	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	·9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	·9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	·9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	·9590	9596	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	·9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	·9685	9689	9694	9699	9703	9708	9713	9117	9722	9727	0	1	1	2	2	3	3	4	4
94	·9731	9736	9741	9345	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	·9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	·9823	9827	9832	9836	9841	9845	9850	9854	9359	9863	0	1	1	2	2	3	3	4	4
97	·9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	·9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	·9956	9961	9965	9969	9974	9978	9983	9887	9991	9996	0	1	1	2	2	3	3	3	4

TABLE II
ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
·00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
·01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
·02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
·03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
·04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
·05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
·06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
·07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
·08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
·09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
·10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
·11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
·21	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
·13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
·14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
·15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	1	1	1	2		2	2	3	3
·16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
·17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
·18	1514	1717	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
·19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
·20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
·21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
·22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
·23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
·24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
·25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	2	3	3	4
·26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
·27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
·28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
·29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
·30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
·31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
·32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
·33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
·34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
·35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
·36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
·37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
·38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
·39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	5	5
·40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
·41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
·42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
·44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
·45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
·46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
·47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
·48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	5	5	6	6
·49	3090	3097	3105	3112	3119	3126	3133	3141	5148	3155	1	1	2	3	4	4	5	6	6

TABLE II
ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
·50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
·51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
·52	3311	3319	3327	3334	3342	3550	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
·53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
·54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	3	3	4	5	6	6	7
·55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
·56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
·57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
·58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
·59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	3	4	5	6	7	8
·60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
·61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
·62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
·63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
·64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
·65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
·66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
·67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
·68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
·69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
·70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
·71	5129	5140	5152	5165	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
·72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
·73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
·74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
·75	5623	5636	5649	5662	5675	5889	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
·76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
·77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
·78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
·79	6166	6188	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
·80	6310	6324	6339	5353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
·81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
·82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
·83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
·84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
·85	7079	7096	7112	7129	7445	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
·86	7344	7261	7278	7295	7311	7328	7345	7362	7339	7396	2	3	5	7	8	10	12	13	15
·87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
·88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
·89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
·90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
·91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
·92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
·93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
·94	8710	8730	8750	8770	8790	3810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
·95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
·96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
·97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
·98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
·99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20

TABLE II
POWERS, ROOTS AND RECIPROCALS

n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$	$\sqrt{10n}$	$\sqrt[3]{10n}$	$\sqrt[3]{100n}$	$\frac{1}{n}$
1	1	1	1	1	3·162	2·154	4·642	1
2	4	8	1·414	1·260	4·772	2·714	5·848	·5000
3	9	27	1·732	1·442	5·477	3·107	6·694	·3333
4	16	64	2	1·587	6·325	3·420	7·368	·2500
5	25	125	2·236	1·710	7·671	3·684	7·937	·2000
6	36	216	2·449	1·817	7·745	3·915	8·434	·1667
7	49	343	2·646	1·913	8·361	4·121	8·879	·1429
8	64	512	2·828	2·000	8·944	4·309	9·283	·1250
9	81	729	3·000	2·080	9·487	4·481	9·655	·1111
10	100	1000	3·162	2·154	10·0	4·642	10·000	·1000
11	121	1331	3·317	2·224	10·488	4·791	10·323	·09091
12	144	1728	3·464	2·289	10·954	4·932	10·627	·08333
13	169	2197	3·606	2·351	11·402	5·066	10·914	·07692
14	196	2744	3·742	2·410	11·832	5·192	11·187	·07143
15	225	3375	3·873	2·466	12·247	5·313	11·447	·06667
16	250	4096	4·000	2·520	12·649	5·429	11·696	·06250
17	289	4913	4·123	2·571	13·038	5·540	11·935	·05882
18	324	5832	4·243	2·621	13·416	5·646	12·164	·05556
19	361	6859	4·359	2·568	13·784	5·749	12·386	·05263
20	400	8000	4·472	2·714	14·142	5·848	12·599	·0500
21	441	9261	4·583	2·759	14·491	5·944	12·806	·04762
22	484	10648	4·690	2·802	14·832	6·037	13·606	·04545
23	529	12167	4·796	2·844	15·166	6·127	13·200	·04167
24	576	13824	4·899	2·884	15·492	6·214	13·389	·04167
25	625	15625	5·000	2·924	15·811	6·300	13·572	·0400
26	676	17576	5·099	2·962	16·125	6·383	13·751	·03846
27	729	19683	5·196	3·000	16·432	6·463	13·925	·03704
28	784	21952	5·292	3·037	16·733	6·542	14·095	·03571
29	841	24389	5·385	3·072	17·029	6·619	14·260	·03448
30	900	27000	5·477	3·107	17·321	6·694	14·422	·03333
31	961	29791	5·568	3·141	17·607	6·768	14·581	·03226
32	1024	32768	5·657	3·175	17·889	6·9840	17·736	·03125
33	1089	35937	5·745	3·208	18·166	6·910	14·888	·03030
34	1156	39304	5·831	3·240	18·439	6·980	15·037	·02941
35	1225	42875	5·916	3·271	18·708	7·047	15·183	·02857
36	1296	46656	6·000	3·302	18·974	7·114	15·326	·02778
37	1369	50653	6·083	3·332	19·235	7·179	15·467	·02703
38	1444	54872	6·164	3·362	19·494	7·243	15·605	·02632
39	1521	59319	6·245	3·391	19·748	7·306	15·741	·02564
40	1600	64000	6·325	3·420	20·00	7·368	15·874	·0250
41	1681	68921	6·403	3·448	20·248	7·429	16·005	·02439
42	1764	74088	6·481	3·476	20·494	7·489	16·134	·02381
43	1849	79507	6·557	3·503	20·736	7·548	16·261	·02326
44	1936	85184	6·633	3·530	20·976	7·606	16·386	·02273
45	2025	91125	6·708	3·557	21·213	7·663	16·510	·02222
46	2116	97336	6·782	3·583	21·448	7·719	16·631	·02174
47	2209	103823	6·856	3·609	21·679	7·775	16·751	·02128
48	2304	110592	6·928	3·634	21·909	7·830	16·869	·02083
49	2401	117649	7·000	3·659	22·136	7·884	16·985	·02041
50	2500	125000	7·071	3·684	22·361	7·937	17·100	·020

TABLE III
POWERS, ROOTS AND RECIPROCALS

n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$	$\sqrt{10n}$	$\sqrt[3]{10n}$	$\sqrt[3]{100n}$	$\frac{1}{n}$
51	2601	132651	7·141	3·708	22·583	7·990	17·213	·01961
52	2704	140608	7·211	3·733	22·804	8·041	17·325	·01923
53	2809	148877	7·280	3·756	23·022	8·093	17·435	·01887
54	2916	157464	7·348	3·780	23·238	8·143	17·544	·01852
55	3025	166375	7·416	3·803	23·452	8·193	17·652	·01818
56	3136	175616	7·483	3·832	23·664	8·243	17·758	·01786
57	3249	185193	7·550	3·849	23·875	8·291	17·863	·01754
58	3364	195112	7·616	3·871	24·083	8·340	17·967	·01724
59	3481	205379	7·681	3·893	24·290	8·387	18·070	·01695
60	3600	216000	7·746	3·915	24·495	8·334	18·171	·01667
61	3721	226981	7·810	3·936	24·698	8·481	18·272	·01639
62	3844	238328	7·874	3·958	24·900	8·527	18·371	·01613
63	3969	250047	7·937	3·979	25·100	8·573	18·469	·01587
64	4096	262144	8·000	4·000	25·298	8·618	18·566	·01562
65	4225	274625	8·062	4·021	25·495	8·662	18·663	·01538
68	4356	287496	8·124	4·041	25·690	8·707	18·758	·01515
67	4489	300763	8·185	4·062	25·884	8·750	18·852	·01493
68	4624	314432	8·246	4·082	26·077	8·794	18·945	·01471
69	4761	328509	8·307	4·102	26·268	8·837	19·038	·01449
70	4900	343000	8·367	4·121	26·458	8·879	19·129	·01429
71	5041	357911	8·426	4·141	26·646	8·921	19·220	·01408
72	5184	373248	8·485	4·160	26·833	8·963	19·310	·01389
73	5329	389017	8·544	4·179	27·019	9·004	19·399	·01370
74	5476	405224	8·602	4·198	27·203	9·045	19·487	·01351
75	5625	421875	8·660	4·217	27·38	9·086	19·574	·01333
76	5776	438976	8·718	4·236	27·568	9·126	19·661	·01316
77	5929	456533	8·775	4·254	27·740	9·166	19·747	·01299
78	6084	474552	8·832	4·273	27·928	9·205	19·832	·01282
79	6241	493039	8·888	4·291	28·107	9·244	19·916	·01266
80	6400	512000	8·944	4·309	28·284	9·283	20·000	·01250
81	6561	531441	9·000	4·327	28·460	9·322	20·083	·01235
82	6724	551368	9·055	4·344	28·636	9·360	20·165	·01220
83	6889	571787	9·110	4·362	28·810	9·398	20·247	·01205
84	7056	592704	9·165	4·380	28·983	9·435	20·328	·01190
85	7225	614125	9·220	4·397	29·155	9·473	20·408	·01176
86	7396	636056	9·274	4·414	29·326	9·510	20·488	·01163
87	7569	658503	9·327	4·431	29·496	9·546	20·507	·01149
88	7744	681472	9·381	4·448	29·665	9·583	20·646	·01136
89	7921	704969	7·434	4·465	29·833	9·619	20·224	·01124
90	8100	729000	9·487	4·487	30·000	9·655	20·801	·01111
91	8281	753571	9·539	4·498	30·166	9·691	20·878	·01099
92	8464	775688	9·592	4·514	30·332	9·726	20·954	·01087
93	8649	804357	9·644	4·531	30·496	9·761	21·029	·01075
94	8830	830584	9·695	4·547	30·659	9·796	21·105	·01064
95	9023	857375	9·747	4·563	30·822	9·830	21·179	·01053
96	9216	884736	9·798	4·579	30·984	9·865	21·253	·01042
97	9409	912673	9·849	4·595	31·145	9·899	21·327	·01031
98	9604	941192	9·899	4·610	31·305	9·933	21·400	·01020
99	9801	970299	9·900	4·626	31·464	9·967	21·472	·01010
100	10000	1000000	10·000	4·642	31·623	10·000	21·544	·01000

TABLE IV

AREAS UNDER NORMAL CURVE

Normal probability curve is given by

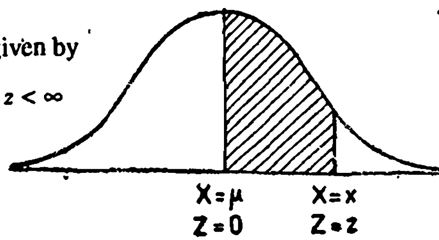
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \quad -\infty < x < \infty$$

and standard normal probability curve is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), \; -\infty < z < \infty$$

where $Z = \frac{X - E(X)}{\sigma_X} \sim N(0, 1)$

Areas under Nomal Curve



The following table gives the shaded area in the diagram *viz.*. $P(0<Z<z)$ for different values of z.

TABLE OF AREAS

↓Z→	0	1	2	3	4	5	6	7	8	9
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0759
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3655	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4678	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4959	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.1960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

TABLE V

ORDINATES OF THE NORMAL PROBABILITY CURVE

The following table gives the ordinates of the standard normal probability curve, *i.e.*, it gives the value of

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2} z^2\right), \quad -\infty < z < \infty$$

for different values of z, where

$$Z = \frac{X - E(X)}{\sigma_X} = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Obviously $\phi(-z) = \phi(z)$.

Z	·00	·01	·02	·03	·04	·05	·06	·07	·08	·09
0·0	·3989	·3989	·3989	·3988	·3986	·3984	·3982	·3980	·3977	·3973
0·1	·3970	·3965	·3961	·3956	·3951	·3945	·3939	·3932	·3925	·3918
0·2	·3910	·3902	·3894	·3885	·3876	·3867	·3857	·3847	·3836	·3825
0·3	·3814	·3802	·3790	·8778	·3765	·3752	·3739	·3725	·3712	·3697
0·4	·3683	·3668	·3653	·3637	·3621	·3605	·3589	·3572	·3555	·3538
0·5	·3521	·3503	·3485	·3467	·3448	·3429	·3410	·3391	·3372	·3352
0·6	·3332	·3312	·3292	·3271	·3251	·3230	·3209	·3187	·3166	·3144
0·7	·3123	·3101	·3079	·3056	·3034	·3011	·2989	·2966	·2943	·2920
0·8	·2897	·2874	·2850	·2827	·2803	·2780	·2756	·2732	·2709	·2685
0·9	·2661	·2637	·2613	·2589	·2565	·2541	·2516	·2492	·2468	·2444
1·0	·2420	·2396	·2371	·2347	·2323	·2299	·2275	·2251	·2227	·2203
1·1	·2179	·2155	·2131	·2107	·2083	·2059	·2036	·2012	·1989	·1965
1·2	·1942	·1919	·1895	·1872	·1849	·1826	·1804	·1781	·1758	·1736
1·3	·1714	·1691	·1669	·1647	·1626	·1604	·1582	·1561	·1539	·1518
1·4	·1497	·1476	·1456	·1435	·1415	·1394	·1374	·1354	·1334	·1315
1·5	·1295	·1276	·1257	·1238	·1219	·1200	·1182	·1163	·1145	·1127
1·6	·1109	·1092	·1074	·1057	·1040	·1023	·1006	·0989	·0973	·0957
1·7	·0940	·0925	·0909	·0893	·0878	·0863	·0848	·0833	·0818	·0804
1·8	·0790	·0775	·0761	·0748	·0734	·0721	·0707	·0694	·0681	·0669
1·9	·0656	·0644	·0632	·0620	·0608	·0596	·0584	·0573	·0562	·0551
2·0	·0540	·0529	·0519	·0508	·0498	·0488	·0478	·0468	·0459	·0449
2·1	·0440	·0431	·0422	·0413	·0404	·0396	·0387	·0379	·0371	·0363
2·2	·0355	·0347	·0339	·0332	·0325	·0317	·0310	·0303	·0297	·0290
2·3	·0283	·0277	·0270	·0264	·0258	·0252	·0246	·0241	·0235	·0229
2·4	·0224	·0219	·0213	·0208	·0203	·0198	·0194	·0189	·0184	·0180
2·5	·0175	·0171	·0167	·0163	·0158	·0154	·0151	·0147	·0143	·0139
2·6	·0136	·0132	·0129	·0126	·0122	·0119	·0116	·0113	·0110	·0107
2·7	·0104	·0101	·0099	·0096	·0093	·0091	·0088	·0086	·0084	·0081
2·8	·0079	·0077	·0075	·0073	·0071	·0069	·0067	·0065	·0063	·0061
2·9	·0060	·0058	·0056	·0055	·0053	·0051	·0050	·0048	·0047	·0046
3·0	·0044	·0043	·0042	·0040	·0039	·0038	·0037	·0036	·0035	·0034
3·1	·0033	·0032	·0031	·0030	·0029	·0028	·0027	·0026	·0025	·0025
3·2	·0024	·0023	·0022	·0022	·0021	·0020	·0020	·0019	·0018	·0018
3·3	·0017	·0017	·0016	·0016	·0015	·0015	·0014	·0014	·0013	·0013
3·4	·0012	·0012	·0012	·0011	·0011	·0010	·0010	·0010	·0009	·0009
3·5	·0009	·0008	·0008	·0008	·0008	·0007	·0007	·0007	·0007	·0006
3·6	·0006	·0006	·0006	·0005	·0005	·0005	·0005	·0005	·0005	·0004
3·7	·0004	·0004	·0004	·0004	·0004	·0004	·0003	·0003	·0003	·0003
3·8	·0003	·0003	·0003	·0003	·0003	·0002	·0002	·0002	·0002	·0002
3·9	·0002	·0002	·0002	·0002	·0002	·0002	·0002	·0002	·0001	·0001

TABLE VI

SIGNIFICANT VALUES $\chi^2(\alpha)$ OF CHI-SQUARE DISTRIBUTION (RIGHT TAIL AREAS FOR GIVEN PROBABILITY α, where

$$P = P_r\,(\chi^2 > \chi^2(\alpha)) = \alpha$$

AND υ IS DEGREES OF FREEDOM (d.f.)

Degree of freedom (υ)	*Probability (Level of significance)* 0 = ·99	*0·95*	*0·50*	*0·10*	*0·05*	*0·02*	*0·01*
1	·000157	·00393	·455	2·706	3·841	5·214	6·635
2	·0201	·103	1·386	4·605	5·991	7·824	9·210
3	·115	·352	2·366	6·251	7·815	9·837	11·341
4	·297	·711	3·357	7·779	9·488	11·668	13·277
5	·554	1·145	4·351	9·236	11·070	13·388	15·086
6	·872	1·635	5·348	10·645	12·592	15·033	16·812
7	1·239	2·167	6·346	12·017	14·067	16·622	18·475
8	1·646	2·733	7·344	13·362	15·507	18·168	20·090
9	2·088	3·325	8·343	14·684	16·919	19·679	21·666
10	2·558	3·940	9·340	15·987	18·307	21·161	23·209
11	3·053	4·575	10·341	17·275	19·675	22·618	24·725
12	3·571	5·226	11·340	18·549	21·026	24·054	26·217
13	4·107	5·892	12·340	19·812	22·362	25·472	27·688
14	4·660	6·571	13·339	21·064	23·685	26·873	29·141
15	4·229	7·261	14·339	22·307	24·996	28·259	30·578
16	5·812	7·962	15·338	23·542	26·296	29·633	32·000
17	6·408	8·672	16·338	24·769	27·587	30·995	33·409
18	7·015	9·390	17·338	25·989	28·869	32·346	34·805
19	7·633	10·117	18·338	27·204	30·144	33·687	36·191
20	8·260	10·851	19·337	28·412	31·410	35·020	37·566
21	8·897	11·591	20·337	29·615	32·671	36·343	38·932
22	9·542	12·338	21·337	30·813	33·924	37·659	40·289
23	10·196	13·091	22·337	32·007	35·172	38·968	41·638
24	10·856	13·848	23·337	32·196	36·415	40·270	42·980
25	11·524	14·611	24·337	34·382	37·652	41·566	44·314
26	12·198	15·379	25·336	35·363	38·885	41·856	45·642
27	12·879	16·151	26·336	36·741	40·113	44·140	46·963
28	13·565	16·928	27·336	37·916	41·337	45·419	48·278
29	14·256	17·708	28·336	39·087	42·557	46·693	49·588
30	14·953	18·493	29·336	40·256	43·773	47·962	50·892

Note. For degrees of freedom (υ) greater than 30, the quantity $\sqrt{2\chi^2} - \sqrt{2\upsilon - 1}$ may be used as a normal variate with unit variance.

TABLE VII

SIGNIFICANT VALUES $t_\nu(\alpha)$ OF t-DISTRIBUTION (TWO TAIL AREAS)

$$P\left[|t| > t_\nu(\alpha)\right] = \alpha$$

d.f. (ν)	*Probability (Level of Significance)* 0·50	0·10	0·05	0·02	0·01	0·001
1	1·00	6·31	12·71	31·82	63·66	636·62
2	0·82	0·92	4·30	6·97	6·93	31·60
3	0·77	2·35	3·18	4·54	5·84	12·94
4	0·74	2·13	2·78	3·75	4·60	8·61
5	0·73	2·02	2·57	3·37	4·03	6·86
6	0·72	1·94	2·45	3·14	3·71	5·96
7	0·71	1·90	2·37	3·00	3·50	5·41
8	0·71	1·80	2·31	2·90	3·36	5·04
9	0·70	1·83	2·26	2·82	3·25	4·78
10	0·70	1·81	2·23	2·76	3·17	4·59
11	0·70	1·80	2·20	2·72	3·11	4·44
12	0·70	1·78	2·18	2·68	3·06	4·32
13	0·69	1·77	2·16	2·05	3·01	4·22
14	0·69	1·76	2·15	2·62	2·98	4·14
15	0·69	1·75	2·13	2·60	2·95	4·07
16	0·69	1·75	2·12	2·58	2·92	4·02
17	0·69	1·74	2·11	2·57	2·90	3·97
18	0·69	1·73	2·10	2·55	2·88	3·92
19	0·69	1·73	2·09	2·54	2·86	3·88
20	0·69	1·73	2·09	2·53	2·85	3·85
21	0·69	1·72	2·08	2·52	2·83	3·83
22	0·69	1·72	2·07	2·51	2·82	3·79
23	0·69	1·71	2·07	2·50	2·81	3·77
24	0·69	1·71	2·06	2·49	2·80	3·75
25	0·68	1·71	2·06	2·49	2·79	3·73
26	0·68	1·71	2·06	2·48	2·78	3·71
27	0·68	1·70	2·05	2·47	2·77	3·69
28	0·68	1·70	2·05	2·47	2·76	3·67
29	0·68	1·70	2·05	2·46	2·76	3·66
30	0·68	1·70	2·04	2·46	2·75	3·65
∞	0·67	1·65	1·96	2·33	2·58	3·29

TABLE VIII
SIGNIFICANT VALUES OF THE VARIANCE-RATIO F-DISTRIBUTION (RIGHT TAIL AREAS)
5 PER CENT POINTS

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	161·4	199·5	215·7	224·6	230·2	234·0	238·9	243·9	249·0	254·3
2	18·51	19·00	19·16	19·25	19·30	19·35	19·37	19·41	19·45	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·84	8·74	8·64	8·55
4	7·71	6·94	6·59	6·39	6·26	6·16	6·04	5·91	5·77	5·65
5	6·61	5·79	5·41	5·19	5·05	4·95	4·82	4·68	4·53	4·96
6	5·99	5·14	4·76	4·53	4·39	4·28	4·15	4·00	3·84	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·78	3·57	3·41	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·44	3·28	3·12	2·93
9	5·12	4·26	3·865	3·63	3·48	3·37	3·23	3·07	2·90	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·07	2·91	2·74	2·54
11	4·84	3·98	3·59	3·365	3·20	3·09	2·95	2·79	2·61	2·40
12	4·75	3·88	4·49	3·26	3·11	3·00	2·85	2·69	2·50	2·30
13	4·67	3·80	5·41	3·18	3·02	2·92	2·77	2·60	2·42	2·21
14	4·60	3·74	3·54	3·11	2·96	2·85	2·70	2·53	2·35	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·64	2·48	2·29	2·07
16	4·49	3·63	3·[illegible]4	3·01	2·85	2·74	2·59	2·42	2·24	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·55	2·38	2·19	1·96
18	4·41	3·55	3·96	2·93	2·77	2·66	2·51	2·34	2·15	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·48	2·31	2·11	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·45	2·28	2·08	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·42	2·25	2·05	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·40	2·23	2·03	1·76
23	4·28	3·42	3·03	2·80	2·64	2·53	2·38	2·20	2·00	1·76
24	4·26	4·40	3·01	2·78	2·62	2·51	2·36	2·18	1·98	1·73
25	4·24	3·38	2·99	2·76	2·60	2·49	2·34	2·16	1·96	1·71
26	4·22	3·37	2·98	2·74	2·59	2·47	2·32	2·15	1·95	1·60
27	4·21	3·35	2·96	2·73	2·57	2·46	2·30	2·13	1·93	1·67
28	4·20	3·34	2·95	2·71	2·56	2·44	2·29	2·12	1·91	1·65
29	4·18	3·33	2·93	2·70	2·54	2·43	2·28	2·10	1·90	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·27	2·09	1·89	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·18	2·00	1·79	1·51
60	4·00	3·15	2·76	2·52	2·37	2·25	2·10	1·92	1·70	1·30
120	3·92	3·87	2·68	2·45	2·29	2·17	2·02	1·83	1·62	1·25
240	3·84	2·99	2·60	2·37	2·21	2·09	1·94	1·75	1·52	1·00

Index

H

I

J

K

L

M

N