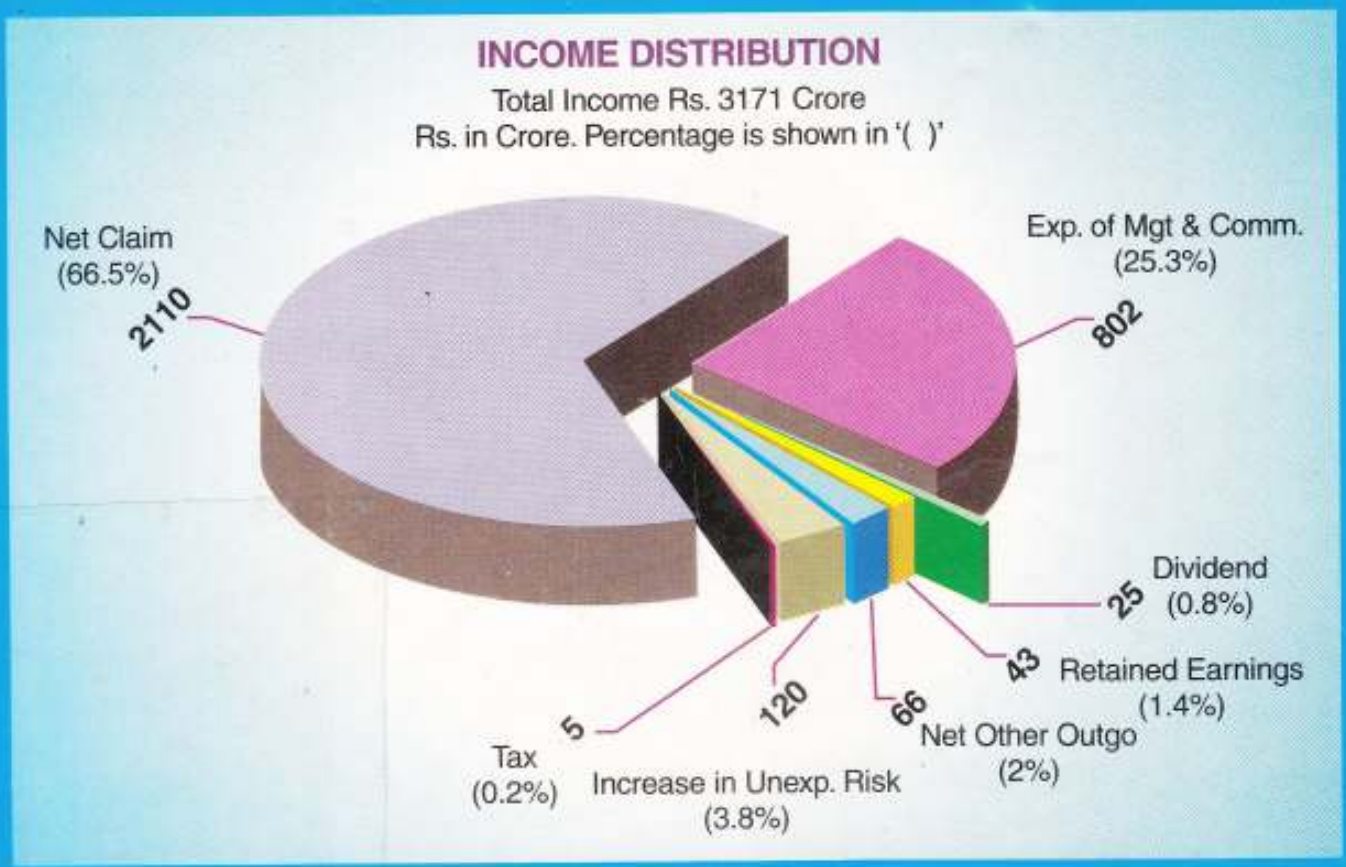
S.P. GUPTA ■ M.P. GUPTA

BUSINESS STATISTICS



BUSINESS STATISTICS

BUSINESS STATISTICS

Dr. S.P. GUPTA

M.Com., Ph.D. (Delhi)
Formerly, Head & Dean,
Faculty of Management Studies,
University of Delhi, Delhi

Dr. M.P. GUPTA

M.A. (Eco.); M.A. (O.R.); Ph.D.
Formerly, Head & Dean,
Faculty of Management Studies,
University of Delhi, Delhi.

Sixteenth Enlarged Edition

Sultan Chand & Sons

Educational Publishers

New Delhi

Other Books by Dr. S.P. Gupta

- Statistical Methods (All India) 39th Edition
- Statistical Methods for Professional Education Course
- Elementary Statistical Methods for B.Com.
- Statistical Methods for B.Com. (Hons.), B.A. (Hons.) Econ.
- Business Statistics & Operations Research
- Quantitative Methods for B.Com. I year, A.P. State
- Quantitative Methods for B.Com. II year, A.P. State
- Objective Type Questions in Statistics
- Statistical Methods (Hindi Edn.)

Fifteenth Edition 2008
Sixteenth Edition 2010

ISBN : 978-81-8054-641-9

Price : Rs. 325.00

Published by:
SULTAN CHAND & SONS
23, Daryaganj, New Delhi-110002
Phones: 23243183, 23247051, 23266105, 23277843, 20262215
Fax: 011-2326-6357

Printed at: New A.S. Offset, 4/203 Lalita Park, Laxmi Nagar, Delhi-110092

Preface

To the Sixteenth Edition

The text is written with the basic object of introducing students of business administration to the Statistical concepts that help in decision-making. An attempt has been made to present explanation in such a way that the underlying statistical theory is fully exposed and the relation between theory and application thoroughly understood.

The book is essentially non-mathematical in character and an attempt has been made to illustrate the application of statistical techniques with the help of business data. Various types of study material are given at the end of each chapter to aid the students in applying the principles discussed in the text. The object is to develop the faculty of thinking amongst the students and to develop the skill of performing the calculations needed for various methods of analysis.

We are greatly inspired by the very good response from a large number of readers of Business Statistics in India and abroad.

Some special features of this edition are :

- The entire text of the SIXTEENTH edition has been thoroughly revised. In particular, the chapters • Correlation Analysis • Probability • Probability Distributions • Statistical Decision Theory • Small Sampling Theory • Chi-Square Test need special mention.
- Looking to the trend of questions of several universities all over India, short answer questions of 1 mark and 4 marks have been added in each chapter to enhance the value of the book.
- Every effort has been made to reduce to the minimum the printing and calculation mistakes.

We gratefully acknowledge all suggestions received to enhance further the value of the text. The suggestions have been incorporated wherever possible.

We are grateful to Prof. Surendra Pradhan of Hayward University, California, Mr. Rajeev Gulhar Applied Materials, USA., Dr. Sarika Gulhar and Mr. Sameer Gupta for help in the revision of this edition.

We sincerely believe that the road to improvement is never ending. Hence, we shall look forward to and gratefully acknowledge all suggestions received.

1st July, 2010

S.P. GUPTA
M.P. GUPTA

Acknowledgements

The authors gratefully acknowledge the inspiration, encouragement, guidance, help and valuable suggestions received from the following well-wishers.

Prof. Abad Ahmad, Formerly PVC, University of Delhi.

Prof. B.S. Sharma, Formerly Vice Chancellor, Kota Open University, Kota.

Dr. Ganesh Mani, Head, Dept. of Cardiology, Delhi Heart & Lung Institute, ND.

Dr. A.B. Ghosh, FRCS, North End Medicare Centre, Delhi.

Prof. B.P.Singh, Former Dean, Faculty of Commerce & Business, University of Delhi.

Prof. Thomas Gladwin, New York University, Graduate, School of Business Administration, USA.

Prof. R.N. Goyale, Former Head, Dept. of Business Studies, University of Delhi.

Mr. Rajeev Gulhar, Sr. Systems Manager, Applied Materials, USA.

Ms. Sarika Gulhar, HRD Manager, CASCADE Promotion Corporation, California.

Mr. Sameer Gupta, Wipro Technologies, USA.

Prof. J.D. Aggarwal, Executive Director, Indian Institute of Finance, Delhi.

Prof. Y.P. Singh, Former Head, Dept. of Business Studies, University of Delhi.

Prof. S.K. Gupta, Florida International University, Miami, USA.

Prof. H.B. Singh, University of Delhi, Delhi.

Prof. N.S. Bisht, Head, Dept. of Commerce, Kumaun University, Nainital.

Prof. N.L. Dhamija, MDI, Gurgaon.

Mr. G.P. Gupta, Former Chairman, IDBI.

Prof. K.L. Krishna, Delhi School of Economics, University of Delhi.

Prof. T.C. Majupuria, Tribhuvan University, Kathmandu, Nepal.

Prof. B.N. Nagnur, Dept. of Statistics, Karnataka University, Dharwar.

Prof. Gabor Parniczky, Karl Marx University, Budapest.

Prof. Surendra Pradhan, California State University, California.

Late Mr. Prakash Chand, Sultan Chand & Sons, New Delhi.

Prof. Nageshwar Rao, Director, Pt. Jawaharlal Nehru Institute of Management, Ujjain.

Prof. J.V Prabhakar Rao, Dept. of Commerce & Management Studies, Andhra University, A.P.

Dr. Govardhan Reddy, Head, Dept. of Commerce, Osmania University, Hyderabad.

Prof. Y.P. Sabharwal, Dept. of Mathematical Statistics, Ramjas College, University of Delhi.

Dr. N.C. Goel, Pitampura, Delhi.

Prof. Kanwar Sen, Dept. of Statistics, University of Delhi.

Prof. Fayyaz Ahmad, Dean, Faculty of Commerce & Management, University of Kashmir.

Mrs. & Dr. Vijay Kansal, Pitampura, Delhi.

Dr. Subhash Talwar, Pitampura, Delhi.

Brief Contents

Contents

CALCULATION OF REGRESSION EQUATION

Salesmen	Test Score X	$(X-\bar{X})$ x	x^2	Sales Y	$(Y-\bar{Y})$ y	y^2	xy
1	40	–20	400	2.5	–1.5	2.25	+30
2	70	+10	100	6.0	+2.0	4.00	+20
3	50	–10	100	4.0	0	0	0
4	60	0	0	5.0	1.0	1.00	0
5	80	+20	400	4.0	0	0	0
6	50	–10	100	2.5	–1.5	2.25	+15
7	90	+30	900	5.5	+1.5	2.25	+45
8	40	–20	400	3.0	–1.0	1.00	+20
9	60	0	0	4.5	+0.5	0.25	0
10	60	0	0	3.0	–1.0	1.00	0
$N=10$	$\Sigma X=600$	$\Sigma x=0$	$\Sigma x^2=2,400$	$\Sigma Y=40$	$\Sigma y=0$	$\Sigma y^2=14$	$\Sigma xy=$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{600}{10} = 60; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{10} = 4$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{130}{2,400} = 0.054$$

The regression equation of sales and test scores is given as :

$$Y - 4 = 0.054\,(X - 60)$$

$$Y_s = 0.76 + 0.054\,X$$

When X is 100, Y would be

$$Y = 0.76 + 0.054\,(100) = 6.16.$$

Thus the most probable weekly sales volume if salesman makes a score of 100 is 6.16 thousand rupees.

Deviations taken from Assumed Means

When actual means of X and Y variables are in fractions, the calculations can be simplified by using the deviations from the assumed mean. The value of b, *i.e.*, the regression coefficient, will be calculated as follows :

Regression equation of X on Y : $(X-\bar{X}) = b_{xy}(Y-\bar{Y})$

where $$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

Regression equation of Y on X : $(Y-\bar{Y}) = b_{yx}(X-\bar{X})$

where $$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

Once the values of b_{xy} and b_{yx} are determined in the above manner, the regression equations can be obtained very easily.

Illustration. 4. A company wants to assess the impact of R & D expenditure on its annual profit. The following data presents the information for the last eight years :

Years	2010	2009	2008	2007	2006	2005	2004	2003
R & D expenditure (Rs. '000)	9	7	5	10	4	5	3	2
Annual Profit (Rs. '000)	45	42	41	60	30	34	25	20

Estimate the regression equation and predict the annual profit for 2009 for an allocated sum of Rs. 100,000 as R & D expenditure.

Solution. Let R & D expenditure be denoted by X and annual profit by Y.

Operations Research

Techniques for Management

V.K. KAPOOR

Co-author of Business Mathematics, Statistics, Fundamentals of Mathematical Statistics, etc.

7th Revised Edn Pages *xviii* + 1077 Chapters 20 ISBN 81-7014-828-6

The book completely covers the syllabi of M.B.A., M.M.S. and M.Com., C.A. Engineering, Computer Science & Services Exams of all Indian Universities. This well-organised and profusely illustrated book presents updated account of the Operations Research Techniques.

Special Features

It is lucid and practical in approach. Wide variety of carefully selected, adapted and specially designed problems with complete solutions and detailed workings. 221 Worked examples are expertly woven into the text. Useful sets of 740 problems as exercises are given.

Contents

Introduction to Operations Research • Linear Programming : Graphic Method • Linear Programming : Simplex Method • Linear Programming : Duality • Transportation Problems • Assignment Problems • Sequencing Problems • Replacement Decisions • Queuing Theory • Decision Theory and Decision Analysis • Game Theory • Inventory Management • Statistical Quality Control • Investment Project Management • PERT & CPM • Simulation • Work Study • Value Analysis • Markov Analysis • Goal Integer and Dynamic Programming • Appendix : Hints and Answers to Selected Questions.

Problems and Solutions in Operations Research

V.K. KAPOOR

Fourth Rev. Edn. 800 Solved Problems Pp. *xii* + 835 ISBN 81-7014-605-4

Salient Features

- The book fully meets the course requirements of Management and Commerce students.
- Working rules, aid to memory, short-cuts, alternative methods are special attractions of the book.
- Ideal book for the students involved in independent study.

Linear Programming and Decision Making

Dr. A.S. NARAG

Ex Dean, Faculty of Management Studies, University of Delhi, Delhi

Fourth & Revised reprint Edn. 2005 Pp. *x* + 242 Chapters 8 ISBN 81-7014-851-0

It has been written for use primarily in Management, Commerce, Economics, Engineering and other professional disciplines.

This book deals with Linear Programming and its extensions like transportation model, assignment models, etc. which are very well explained with illustrative examples which closely resemble real world problems.

Strategic Planning and Management

Dr. P.K. GHOSH, M.A., Ph.D.
Formerly Professor of Commerce, University of Delhi
Delhi School of Economics, Delhi

10th Revised Edn. Reprint Pp. *xxiv* + 728 ISBN 81-8054-069-3

It is addressed primarily to the post-graduate students of Indian universities and Institutes of Management. However, the text has been put across in such a manner that Indian executives will also find it stimulating.

Special Features

- The book provides an analytical framework for understanding a total organisation in the complex dynamic environment of today. The analysis is in terms of the multiple decision variables concerning strategic management.
- Updating of environmental configuration like SEBI Guidelines for Book-building, Safety-net or Buy-back arrangement and Government Policy on foreign technology agreement.
- Relevant environment factors of significance to corporate houses in the present-day context.
- Turnaround of Sick Industrial Companies, Corporate Governance, and Universal Inner Structure of Effective Leaders.
- Suggestions on 'Four Routes' to securing strategic advantage and 'Generic Strategies' for sustainable competitive advantage.
- Case studies of corporate response to competitiveness.
- Nine new cases reflecting the current reality and 38 others.

Contents

Entrepreneurial Development in India

Dr. C. B. GUPTA
M.Com., Ph.D.
Reader,
Shri Ram College of Commerce,
University of Delhi, Delhi.

Dr. N.P. SRINIVASAN
M.Com., Ph.D.
IFCI Professor of Commerce,
University of Madras, Madras

5th Edn. Reprint Pp. *xii* + 618 ISBN 81-7014-801-4

Contents

PLEASE CUT ALONG THIS LINE AND MAIL TO US

Feedback Prize Contest

NO ENTRY FEE

We propose to mail our readers a 'Supplement' relevant to the subject-matter of this book or 'A Word about Your Career' or 'Pearls of Wisdom' or 'Secrets of Success' on receipt of your 'Feedback'. Further, you can win a prize too !! For this purpose, please fill this coupon and send it along with your 'Feedback' to us at **M/s Sultan Chand & Sons, 23, Daryaganj, New Delhi-110002,** at an early date. To avoid duplication, please inform what you had received earlier. This is without obligation.

How did you come to know of this book : Recommended by your Teacher/ Friend/Bookseller/Advertisement .....

Date of Purchase .....

Year/Edition of the book purchased by you .....

Month and Year of your next examination .....

Name & Address of the Supplier .....

.....

Name of the Teacher who recommended you this book .....

Name and Address of your School/Institution .....

.....

Your Name .....
(IN BLOCK LETTERS)

Your Residential Address .....
(IN BLOCK LETTERS)

.....

Course for which you are studying .....

Please enclose latest Syllabus/Question paper .....

I bought this book because .....

.....

.....

Feedback

Now You can win a prize too !!

Dear Reader

Reg. ***Business Statistics*** by *S.P. Gupta & M.P. Gupta.*

Has it occurred to you that you can do the students/the future readers a favour by sending your suggestions/comments to improve the book ? In addition, a surprise gift awaits you if you are kind enough to let us have your frank assessment, helpful comments/ specific suggestions in detail about the book on a separate sheet as regards the following :

1. Which topics of your syllabus are inadequately or not discussed in the book from the point of view of your examination ?

   .....
   .....
   .....

2. Is there any factual inaccuracy in the book ? Please specify.

   .....
   .....
   .....

3. What is your assessment of this book as regards the presentation of the subject-matter, expression, precision and price in relation to the other books available on this subject ?

   .....
   .....
   .....

4. Which competing books you regard as better than this ? Please specify their authors and publishers.

   1. .....
   2. .....
   3. .....

5. Any other suggestion/comment you would like to make for the improvement of the book ?

   .....
   .....
   .....

Further, you can win a prize for the best criticism on presentation, contents or quality aspect of this book with useful suggestions for improvement. The prize will be awarded each month and will be in the form of our publications as decided by the Editorial Board.

Please feel free to write to us if you have any problem, complaint or grievance regarding our publications or a bright idea to share. We work for you and your success and your Feedback are valuable to us.

Thanking you,

Yours faithfully

Sultan Chand & Sons

1

Business Statistics — What and Why ?

INTRODUCTION

The word "Statistics" is very popularly used in practice. It conveys a variety of meanings to people many of which are inaccurate or, at the very least, misleading. The average person conceives of 'statistics' as column of figures, zig-zag graphs or tables like statistics of production, consumption, per capita income, imports, exports, crimes, divorce, share prices, etc. Such statistics are quite commonly found in newspapers, journals, reports and one can also hear them on radio, television, classroom lectures, etc. For example, one may find statements like 'The production of foodgrains is expected to increase from 230.78 m. tonnes in 2007-08 to 233.88 m. tonnes in 2008-09; the per capita net national product has increased from Rs.15,881 in 1999-00 to Rs. 40,141 in 2008-09. The Planning Commission has opted for an average economic growth of 9 per cent per annum during the Eleventh Five Year Plan period (2007-2012) from 7.6 per cent in the 10th plan. The population of India for the year 2001 is 102.7 crore.

In addition to meaning numerical facts, 'statistics' also refers to a subject, just as 'mathematics' refers to a subject, as well as symbols, formulae and theorems and 'accounting' refers to principles and methods as well as to accounts, balance sheets and income statements. In this sense, Statistics is a body of methods of obtaining and analysing data in order to base decisions on them. It is a branch of scientific methods used in dealing with phenomena that can be described numerically either by count or by measurement. *Thus, the word statistics refers either to quantitative information or to a method of dealing with quantitative information.*

The methods by which statistical data are analysed are called *statistical methods*, although the term is sometimes used more loosely to cover the subject 'Statistics' as a whole. The mathematical theory which is the basis of these methods is called the *theory of statistics* or *mathematical statistics*. Statistical methods are applicable to a very large number of fields—economics, sociology, anthropology, business, agriculture, psychology, medicine, education—all lean heavily upon statistics. Numerous books have been written on business statistics, agricultural statistics, industrial statistics, medical statistics, educational statistics, psychological statistics and other specific areas of application. It is true, of course, that these diversified fields demand somewhat different and specialized technique in particular problems, yet the fundamental principles that underlie the various methods are identical regardless of the field of application. This will become evident to the reader if he realizes that *statistical methods in general are nothing but a refinement of everyday thinking*. They are specially appropriate for handling data which are subject to variation that cannot be fully controlled by experimental method and for which we can have only a fraction of the totality of observations which may exist.

It should be noted at the very outset that Statistics is usually not studied for its own sake ; rather, it is widely employed as a tool—and a highly valuable one—in the analysis of problems in natural, physical and social sciences. In the latter area, statistics often assumes its greatest importance in the study of economics and business. Statistical methods are used by governmental bodies, private business firms, and research agencies as an indispensable aid in (1) forecasting, (2) controlling, and (3) exploring.

Statistical methods range from the most elementary descriptive devices which may be understood by the common man to those complicated mathematical procedures which can be apprehended only by the expert theoreticians. The purpose of this text is to discuss the fundamental principles and techniques of Business Statistics in simple and easily comprehensible manner without going into the highly mathematical aspects of the subject.

STATISTICS DEFINED

There have been many definitions of the term 'Statistics'—indeed scholarly articles have carefully collected together hundreds of definitions, some have defined Statistics* as statistical data (plural sense) whereas others as statistical methods (singular sense). A few definitions are examined in the following paragraph.

STATISTICAL DATA

Quantitative or numerical information may be found almost everywhere in business, industries, economics and many other areas. It is probably more common to refer to data in quantitative form as *statistical data*. But not all numerical data is statistical and hence it is necessary to examine a few definitions of Statistics to understand the characteristics of statistical data.

Webster defined statistics as "*the classified facts relating the condition of the people in a State especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangements.*"

The above definition is too narrow as it confines the scope of statistics to only such facts and figures which relate to the conditions of the people in a State.

Yule and kendall defined statistics as : "By statistics we mean quantitative data affected to a marked extent by multiplicity of causes."

This definition is less comprehensive than the one given by Prof. Horace Secrist who defined statistics as follows :

"By Statistics we mean *aggregate 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.*"

.This definition clearly points out certain characteristics which numerical data must possess in order that they may be called statistics. They are as follows :

1. *Statistics are aggregate of facts.* Single and isolated figures are not statistics for the simple reason that such figures are unrelated and cannot be compared. For example, a single figure relating to production, sale, birth, death, employment, purchase, accident, etc., cannot be regarded statistics although aggregates of such figures would be called statistics because of their comparability and relationship as part of a common phenomenon.

2. *Statistics are affected to a marked extent by multiplicity of causes.* Generally speaking, facts and figures are affected to a considerable extent by a number of forces operating together. For example, statistics of production of rice are affected by the rainfall, quality of soil, seeds and manure, method of cultivation, etc. It is very difficult to study separately the effect of these forces on the production of rice. The same is true of statistics of prices, imports, exports, sales, profits, etc. In the experimental sciences like Physics and Chemistry it is possible to isolate the effect of various forces on a particular event. Ways and means were also being devised in 'Statistics' for segregating the effects of various forces on an event. However, it has proved to be a difficult task in statistical studies of phenomena which are influenced by a complex variety of factors many of which are not measurable.

---

* "Statistics is the use of data to help the decision maker reach better decisions." —*Philip Enus*

3. *Statistics are numerically expressed.* All statistics are numerical statements of facts, *i.e.*, expressed in numbers. Quantitative statements such as 'The population of India is rapidly increasing' ; or 'The production of wheat is not sufficient ; or 'Textile industry is getting sick' do not consitute statistics. The reason is that such statements are vague and one cannot make out anything from them. On the other hand, the statement that the population of India increased from 846.30 million in 1991 to 1003.24 million in 2001 is a statistical statement.

4. *Statistics are enumerated or estimated according to reasonable standards of accuracy.* Facts and figures about any phenomenon can be derived in two ways, *viz.*, by actual counting and measurement or by estimate. Estimate cannot be as precise and accurate as actual count or measurement. For example, an estimate that two lakh College and University teachers participated in a recent strike does not mean exactly two lakh ; it may be a few hundreds or thousands more or less. On the other hand, if we count the number of employees in an organisation and say that there are 60 employees ; this figure would be 100% accurate. In many cases 100% accuracy of numbers may be difficult to attain. The degree of accuracy desired largely depends upon the nature and object of the enquiry. For example, in measuring heights of persons even inches are important whereas in distance between two places, say, Delhi and Mumbai, even metres can be ignored. Hence, in many statistical studies mathematical accuracy cannot be attained. However, it is important that reasonable standards of accuracy should be attained, otherwise numbers may be altogether misleading.

5. *Statistics are collected in a systematic manner.* Before collecting statistics, a suitable plan of data collection should be prepared and the work carried out in a systematic manner. Data collected in a haphazard manner would very likely lead to fallacious conclusions.

6. *Statistics are collected for a pre-determined purpose.* The purpose of collecting data must be decided in advance. The purpose should be well defined and specific. A general statement of purpose is not enough. For example, if the objective is to collect data on prices it would not serve any useful purpose unless one knows whether he wants to collect data on wholesale or retail prices and what are the relevant commodities in view.

7. *Statistics should be placed in relation to each other.* If numerical facts are to be called statistics, they should be comparable. Statistical data are often compared period-wise called chronological comparison or regionwise called geographical comparison. For instance, the per capita income of India for the year 2009-10 may be compared with that of earlier years of with the per capita income of other countries, say U.S.A., U.K., Japan, etc. Valid comparisons can be made only if the data are homogeneous, *i.e.*, relate to the same phenomenon or subject and only likes are compared with likes. It would be meaningless to compare the sales of a small shop with the sales of a big departmental store.

In the absence of the above characteristics numerical data cannot be called statistical and hence, *"all statistics are numerical statements of facts but all numerical statements of facts are not statistics."*

STATISTICAL METHODS

The large volume of numerical information gives rise to the need for systematic methods which can be used to organise, present, analyse and interpret the information effectively. The term statistics has been defined differently by different writers. A few definitions are examined below.

Prof. A.L. Bowley has given three definitions. At one place he says, "*Statistics may be called the science of counting.*" This definition is too narrow because it covers only one aspect of the science, namely, the collection of data. Other aspects like presentation, analysis, interpretation, etc., are completely ignored.

At another place Bowley says, "*Statistics may be called the science of averages.*" This definition also is not satisfactory because averages are only one of the devices used in statistical analysis. The other devices like variation, skewness, correlation, etc., are not at all covered by this definition.

Still another definition given by the same author is "*Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations.*" This definition also in not satisfactory because it confines the scope of statistics only to sociology, *i.e.*, man and his activities. Bowley himself recognized this when he remarked, "Statistics cannot be confined to any one science."

Boddington defines statistics as " *the science of estimates and probabilities.*" This definition also is unacceptable because estimates and probabilities are only a part of statistical methods.

Croxton and Cowden have given a very simple and concise definition of statistics. In their view, "*Statistics may be defined as a science of collection, presentation, analysis and interpretation of numerical data.*" This definition clearly points out four stages in a statistical investigation, namely : (*i*) collection of data, (*ii*) presentation of data, (*iii*) analysis of data and (*iv*) interpretation of data.

However, to the above stages one more stage may be added and that is the organisation of data. Thus, statistics may be defined as *the science of collection, organisation, presentation, analysis and interpretation of numerical data.*

According to the above definition, there are five stages in a statistical investigation :

1. *Collection.* Collection of data constitutes the first step in a statistical investigation. Utmost care must be excercised in collecting data because they form the foundation of statistical analysis. If data are faulty, the conclusions drawn can never be reliable. The data may be available from existing published or unpublished sources or else may be collected by the investigator himself. The firsthand collection of data is one of the most difficult and important tasks faced by a statistician. Therefore, like all scientific pursuits, the investigator must take into account whatever data have already been collected by others. This would save the investigator from foreseable pitfalls, unnecessary labour and duplication of efforts.

2. *Organization.* Data collected from published sources are generally in organized form. However, a large mass of figures that are collected from a survey frequently needs organisation. The first step in organizing a mass of data is *editing.* The collected data must be edited very carefully so that the omissions, inconsistencies, irrelevant answers and wrong computation in the returns from a survey may be corrected or adjusted. After the data have been edited the next step is to *classify* some common characteristics possessed by the items constituting the data. The last step in organization is *tabulation.* The object of tabulation is to arrange the data in columns and rows so that there is absolute clarity in the data presented.

3. *Presentation.* After the data have been collected and organized they are ready for presentation. Data presented in an orderly manner facilitates statistical analysis.

4. *Analysis.* After collection, organization and presentation the next step is that of analysis. A major part of this text is devoted to the methods used in analysing the presented data, mostly in a tabular form. Methods used in analysing the presented data are numerous ranging from simple observation of data to complicated, sophisticated and highly mathematical techniques. However, in this text only the most commonly used methods of statistical analysis are included.

5. *Interpretation.* The last stage in statistical investigation is interpretation, *i.e.*, drawing conclusions from the data collected and analysed. The interpretation of data is a difficult task and necessitates a high degree of skill and experience. If the data that have been analysed are not properly interpreted, the whole object of the investigation may be defeated and fallacious conclusions drawn. Correct interpretation will lead to a valid conclusion of the study and thus can aid in decision-making.

STATISTICS : SCIENCE OR ART

Whether Statistics is a science or an art is often a subject of debate. Science refers to a systematised body of knowledge. it studies cause and effect relationship and attempts to make generalisations in the form of scientific principles or laws. It describes facts objectively and avoids vague judgements as good

Science, in short, is like a lighthouse that gives light to the ships to find out their own way but does not indicate the direction in which they should go. Art, on the other hand, refers to the skill of handling facts so as to achieve a given objective. It is concerned with ways and means of presenting and handling data, making inferences logically and drawing relevant conclusions.

While a century ago there were some misgivings among natural scientists as to whether statistics had the right to be recognised as a distinct science, now almost *all sciences statistical*. This suggests that the design of scientific experiments and the evaluation of their results makes use of principles and practices growing out of the science of statistics. However, statistics as a science is not similiar to exact sciences like Physics, Chemistry, Zoology, etc. This is because statistical phenomena are generally affected by multiplicity of causes which cannot always be measured accurately. In other words, the science of statistics by its very nature is less precise than the natural sciences. It is science only in a limited sense, viz., as a specialised branch of knowledge. More appropriately, statistics may be regarded as a scientific method because it is really a tool which can be used in scientific studies. Wallis and Roberts have rightly remarked that "Statistics is *not a body of substantive knowledge but a body of methods for obtaining knowledge.*"

If science is knowledge, then art is action. Looking from this angle statistics may also be regarded as an art. It involves the application of given methods to obtain facts, derive results, and finally to use them for devising action.

FUNCTIONS OF STATISTICS

1. It presents facts in a definite form.
2. It simplifies mass of figures.
3. It facilitates comparison.
4. It helps in formulating and testing hypothesis.
5. It helps in prediction.
6. It helps in the formulation of suitable policies.

1. *Definiteness.* Numerical expressions are convincing and, therefore, one of the most important functions of statistics is to present general statements in a precise and definite form. Statements or facts conveyed in exact quantitative terms are always more convincing than vague utterances. Statistics present facts in a precise and definite form and thus help proper comprehension of what is stated. Consider, for example, a statement sex ratio (*i.e.,* number of females per 1000 males) is going up in India. The reader would not have a clear idea of the situation from this statement. But if we say the sex ratio has gone up from 926 in 1991 to 933 in 2001, it conveys a definite meaning. Similarly, statement like 'There is a lot of unemployment in India', 'the population of India is growing at a very fast rate', 'the prices of various commodities are rising', 'the number of students seeking admission to professional courses is increasing', etc. hardly convey any worthwhile information as they do not specify the numerical dimensions involved.

2. *Condensation.* Not only does Statistics present facts in a definite form but it also helps in condensing mass of data into a few significant figures. In a way, statistical methods present a meaningful overall information from the mass of data. Thus, it is impossible for one to form a precise idea about the income position of the people of India from a record of individual incomes of the entire population. However, the figure of per capita income can be easily remembered by everyone.

3. *Comparison.* Unless figures are compared with others of the same kind they are often devoid of any meaning. For example, if we say that the production of Maruti Udyog Ltd. has increased considerably shall not be meaningful unless some comparison of figures is made. But the statement there has been an increase from 200 cars a day in Sept. 1988 to more than 4,000 cars a day in Jan. 2009-2010 definitely indicates the increasing trend in production. Maruti Udyog has grown global and shall be exporting to European and other markets.

4. *Formulating and Testing Hypothesis*. Statistical methods are extremely useful in formulating and testing hypothesis and to develop new theories. For example, hypothesis like whether chloromycetin is effective in preventing typhoid, whether the credit squeeze is effective in checking price increase, whether students have benefited from the extra coaching, etc., can be tested by appropriate statistical tools.

5. *Prediction*. Plans and policies of organisations are invariably formulated well in advance of the time of their implementation. A knowledge of future trends is very helpful in framing suitable policies and plans. Statistical methods provide helpful means of forecasting future events. For example, if a businessman has to decide how much he should produce in 2015, he would like to know the expected sales for that year. He may use his subjective judgment and make a guess. However, a better method for him would be to analyse the sales data of the past years or arrange a statistical survey of the market to obtain necessary data for estimating the sales volume for the year 2015.

6. *Formulation of policies*. Statistics provide the basic material for framing suitable policies. For example, it may be necessary to decide how much oil a nation should import in 2015, the decision would depend upon the expected internal production and the likely demand for oil in 2015. In the absence of information regarding the estimated domestic output and demand for oil the decision on imports cannot be made with reasonable accuracy.

Robert W. Burgess has beautifully summed up the functions of statistics as "*The fundamental gospel of statistics is to push back the domain of ignorance, rule of thumb, arbitrary or premature decisions, traditions and dogmatism and to increase the domain in which decisions are made and principles are formulated on the basis of analysed quantitative facts.*"

SCOPE OF STATISTICS

The scope of statistics is so vast and ever-increasing that not only it is difficult to define but also unwise to do so. The use of Statistics has permeated almost every facet of our lives. It is a tool of all sciences indispensable to research and intelligent judgment and has become a recognized discipline in its own right. There is hardly any field whether it be trade, industry or commerce, economics, biology, botany, astronomy, physics, chemistry, education, medicine, sociology, psychology, or technology where statistical tools are not applicable. In fact, the greatest victory of mankind of the 20th century, that of landing of Apollo 11 on the moon, would not have been a success in the absence of statistical help. The applications of statistics are so numerous that it is often remarked "*Statistics is what statisticians do*". Governments, businessmen and individuals collect statistical data required to carry out their activities efficiently and effectively. Let us examine a few fields in which Statistics is applied.

(*i*) Statistics and State

Since ancient times the ruling kings and chiefs have relied heavily on statistics in framing suitable military and fiscal policies. Most of the statistics such as that of crimes, military strength, population, taxes, etc., that were collected by them were a by-product of administrative activity. In recent years the functions of the State have increased tremendously. The concept of a State has changed from that of simply maintaining law and order to that of a welfare State. Statistical data and statistical methods are of great help in promoting human welfare. Statistics today are not exclusively a by-product of administrative activity—the State collects statistics on several problems. These statistics help in framing suitable policies. All Ministers and departments of the Government whether they be Finance, Transport, Defence, Railways, Food, Commerce, Post and Telegraph or Agriculture, depend heavily on factual data for their efficient functioning. For example, the Transport Department cannot solve the problem of transport in Delhi unless it knows how many buses are operating at present, what is the total

requirement and therefore, how many additional buses be added to the existing fleet. Not only during peace times but during days of war also statistics are indispensable. In fact it is impossible to fight a war successfully in the absence of factual data about enemy strength.

Statistics are so significant to the State that the government in most countries is the biggest collector and user of statistical data. Such data is of immense help to many institutions and research scholars who further process it and arrive at useful conclusions which help in decision-making.

(ii) Statistics in Business and Management

The use of statistical methods in the solution of business problems dates almost exclusively to the 20th century. Applications of statistics pervade virtually every area of activity in business and industry such as production, financial analysis, distribution analysis, market research, research and development, manpower planning and accounting. The main focus of this text is to discuss various statistical techniques that are indispensable in analysing and solving business problems and hence the justification for the book to be called *"Business Statistics"*.

With the growing size and ever-increasing competition, the problems of business enterprise are becoming complex. Prior to Industrial Revolution when production was in the handicraft stage, individual business was small and the production was carried out on a very small scale mainly to cater local needs. The management of such a business enterprises was very much different from the present management of a large-scale business. The information needed by the executives was much less extensive than at present. The market was close at hand, the customers were, for a large part, personal friends of the owner of the business and an elaborate analysis of the market was not needed. The businessman just by asking some questions from his customers could find out what they thought of his product and what they wanted to buy. The manager of a business, who was usually also the owner, worked in the shop with his employees. The number of employees used to be very few and the owner knew them personally and, therefore, records of personnel data were not needed. Similarly, production records were not needed because he knew what progress of work was being made daily in the shop. Any facts he needed could be learnt from direct observation; in fact, most of what he required was in his mind. Thus the owner was directly engaged in all the areas of business activity like sale, purchase, production, finance, accounting, etc.

Under the present system where production is carried out on a large-scale, most of the goods are manufactured in anticipation of consumer demand. Producers and consumers are strangers to each other because it is almost impossible for the producer to know personally thousands and lakhs of consumers. The fact about consumer preferences and desires are not so easy to obtain or so simple to understand as in the earlier system. But since production is carried on in anticipation of demand, such information becomes more important than ever before. It is no longer possible for the owner to see how things are going and to remember each and every aspect of the business. It is also difficult for him to know all his employees personally. Hence with the trend towards large organisation, it has become necessary for the executives to rely more and more on elaborate information systems instead of their intuition or mere observations. And it is here that statistical data and statistical methods play a very significant role. Availability of factual data about the operations of the business is as essential as the availability of raw materials to a manufacturing plant or goods to a retail establishment.

Through the aid of statistical reports, the executive can gain a summary picture of current operations which, improves his factual basis for making valid decisions affecting future operations. The following are some major activities of a typical, large and progressive organisation which would indicate how statistics helps in the efficient discharge of various activities.

1. *Marketing.* Statistical analysis are frequently used in providing information for marketing decisions. In the field of marketing, it is necessary first to find out what can be sold and then to evolve a suitable strategy so that goods reach the ultimate consumer. A skilful analysis of data on population, purchasing power, habits of people, competition, transportation cost, etc., should precede any attempt to establish a new market. Often such analysis will present difficulties which must be properly met before actually attempting to place goods in the market. The analysis may reveal that in certain areas where one thought of big market potential, there hardly exists any scope.

In retail stores, wholesale houses and sales departments and manufacturing concerns, statistical records and analysis enable one to determine in advance, at a small cost, much that would be very costly if determined by actual experience.

In building up and maintaining an extensive market, it is important to keep accurate records of its present and potential geographic distribution. Analysis of sales in relation to the distribution of population and purchasing power are especially important in establishing sales territories routing salesman and in order to know where to advertise and where to push sales.

2. *Production.* In the field of production, statistical data and statistical methods play a very important role. The decision about what to produce, how much to produce, when to produce, for whom to produce is based largely on facts analysed statistically.

Statistical tools are also of immense help in quality control, optimum inventory level and in dealing with labour problems, etc. Production manager looks at quality control data to decide when to make adjustments in a manufacturing process.

3. *Finance.* The Financial Managers in discharging their finance function efficiently depend heavily on statistical analysis of facts and figures. Financial forecasts, breakeven analysis and investment decisions under uncertainty are but part of their activities. In the last three decades, sophisticated models dealing with inventories, cash balances and so on have been developed and applied. These models involve application of several statistical concepts. The area of security analysis is also highly quantitative.

4. *Banking.* Banking institutions have found it increasingly necessary to establish departments within their organisations for the purpose of gathering and analysing information, not only regarding their own operations, but on general economic conditions and on every line of business in which they might be directly or indirectly interested. Probably the banks, more than any other individual business, feel the direct effects of the conditions in every type of business and need to be constantly informed as to the trends in every line of activity. Its reserves are influenced by money markets which are not local but which are national or international ; its funds are influenced not only by the business conditions in the immediate vicinity but also by the conditions of business in areas far distant.

In making loans, banks have to be particularly careful that they do not lend too much money when business is dangerously inflated. In almost every period of hard times some banks have failed because they did not correctly analysed the general business situation or the conditions in individual concerns which had over-expanded.

In all the problems mentioned above, the bankers use the objective analysis furnished by statistics and then temper their decisions on the basis of qualitative information.

5. *Investment.* Statistics greatly assists investor in making clear and valued judgement in his investment decision in selecting securities which are safe and which have the best prospects of yielding a good income. Such investigations assist in determining whether to buy, to sell or to do neither. On the basis of these statistical guides investors purchase securities when they are low, hold them for a few years until they are high and then sell them and hold the proceeds until they can again but at low figures. In this way, without any marginal purchases but by buying outright high grade dividend-paying securities, investors have built up substantial fortunes with relatively little risk.

The investment banker is one of the greatest users of statistics—he must accurately distinguish between good and bad securities. To do so, he should not only have a clear understanding of the present situation of the money and security markets and a definite knowledge of the actual conditions in the different industries but also have a fairly clear conception of what will be the most probable future conditions in various industries.

6. *Purchase.* The purchase department in discharging its functions makes use of statistical data to frame suitable purchase policies such as from where to buy, how much to buy, at what time to buy and at what price to buy.

7. *Accounting.* Statistical methods are also employed in accounting. In particular, the auditing functions makes frequent application of statistical sampling and estimation procedures, and the cost account uses regression analysis. The accountant collects data on historical costs in the course of auditing a company's financial records.

8. *Control.* The management control process combines statistical and accounting methods in making the overall budget for the coming year— including sales, material, labour and other costs, and net profits and capital requirements. It usually maintains a standard cost system for controlling costs, and setting prices of products.

9. *Credit.* The credit department performs statistical analysis to determine how much credit to extend to various customers. In the formulation of future credit policy the characteristics of those who have paid and those who have defaulted are kept in mind.

10. *Personnel.* The personnel department frames personnel policies based on facts. It makes statistical studies of wage rates, incentive plans, cost of living, labour turnover rates, employment trends, accident rates, employee grievances, performance appraisal, training programmes, etc. Such studies help the personnel department in the process of manpower planning.

11. *Research and Development.* Many big organisations have research and development departments which are primarily concerned with finding out how existing products can be improved; what new product lines can be added and how the optimal use of resources made. In the absence of factual data it is almost impossible to carry out fruitful research and development programmes.

(*iii*) Statistics and Economics

In the year 1890 Prof. Alfred Marshall, the renowned economist observed that "*Statistics are the straw out of which I, like every other economists, have to make bricks.*" This proves the significance of statistics in economics. Economics is concerned with the generation and distribution of wealth as well as with the complex institutional set-up concerned with the consumption, saving and investment of income. Statistical data and statistical methods are of immense help in the proper understanding of the economic problems and in the formation of economic policies. In fact, these are the tools and appliances of an economists's laboratory. For example, what to produce, how to produce and for whom to produce—these are the questions that need a lot of statistical data in the absence of which it is not possible to arrive at correct decisions. Statistics of production help in adjusting the supply of demand; Statistics of consumption enable us to find out the way in which people of different strata of society spend their income. Such statistics are very helpful in knowing the standard of living and taxable capacity of the people. In the field of exchange we study markets, laws of prices based on supply and demand, cost of production, banking and credit instruments, etc. What shall be the price of a particualr commodity if its supply increases or decreases ? What price should a monopolist charge in order to reap the maximum profit ?— these are questions which can best be answered with the help of the statistics. In fact, statistics are the very foundation stone of the theory of exchange. In distribution,

too, statistics plays a vital role. How the national income is to be calculated and how it is to be distributed, these are the questions which cannot be answered without statistics. In reducing disparities in the distribution of income and wealth, statistics are of immense help. Similarly in solving problems of rising prices, growing population, unemployment, property, etc., one has to rely heavily on statistics. In fact, most of the economic policies would be a leap in the dark in the absence of appropriate statistical information.

Statistical methods help not only in formulation appropriate economic policies but also in evaluating their effect. For example, in order to check the ever-growing population if emphasis has been placed on the family planning methods, one can ascertain statistically the efficacy of such methods in attaining the desired goal. Statistical techniques play such an important role in the field of economics that in 1926, R.A. Fisher complained of "*the painful misapprehension that statistics is a branch of economics.*"

In recent years *econometrics* which comprises the applications of statistical methods to the theoretical economic methods is widely used in economic research. Statistical methods of sampling are useful for collecting the basic data of economic studies. Statistical methodology also indicates the reliability of the data and the significance to be attached to them. The derivation of demand functions, the field in which the applications of econometrics was first made, continue to be of major interest to economists. Similarly, the production functions, cost functions and the consumption functions present many difficult problems in the analysis of which statistical tools are of immense use.

Thus, economists today are no longer satisfied to theorize in abstract terms. Instead they utilize the excellent data now available to build a sound factual foundation for their reasoning. Some of the applications of statistics in economics are as follows :

1. Measures of gross national product and input-output analysis have greatly advanced overall economic knowledge and opened up entirely new fields of study.

2. Financial statistics are basic in the fields of money and banking short-term credit, consumer finance and public finance.

3. Statistical studies of business cycles, long-term growth and seasonal fluctuations serve to expand our knowledge of economic instability and to modify them from time to time.

4. Studies of competition, oligopoly and monopoly require statistical comparisons of market prices, cost and profits of individual firms.

5. Statistical surveys of prices are essential in studying the theories of prices, pricing policy and price trends, as well as their relationship to the general problem of inflation.

6. Operational studies of public utilities require both statistical and legal tools of analysis.

7. Analysis of population, land economics and economic geography are basically statistical in their approach.

8. In solving various economic problems such as poverty, unemployment, disparities in the distribution of income and wealth, statistical data and statistical methods play a vital role.

In fact, the concept of planning so vital for growth of nations would not have been possible in the absence of data and proper statistical analysis there of.

(*iv*) Statistics and Physical Sciences

The physical sciences, especially astronomy, geology and physics were among the fields in which statistical methods were first developed and applied, but until recently these sciences have not shared the 20th century developments of statistics to the same extent as the biological and social sciences. Currently, however, the physical sciences seem to be making increasing use of statistics, especially in astronomy, chemistry, engineering, geology, meteorolgy and certain branches of physics.

(v) Statistics and Natural Sciences

Statistical techniques have proved to be extremely useful in the study of all natural sciences like biology, medicine, meteorology, zoology, botony, etc. For example, in diagnosing the correct disease the doctor had to rely heavily on factual data like temperature of the body, pulse rate, blood pressure, etc. Similarly, in judging the efficacy of a particular drug for curing a certain disease experiments have to be conducted and the success or failure would depend upon the number of people who are cured after using the drug, In botany—the study of plant life—one has to rely heavily on statistics in conducting experiments about the plants, effect of temperature, type of soil, etc. In fact, it is difficult to find any scientific activity where statistical data and statistical methods are not used.

(vi) Statistics and Research

Statistics is indispensable in research work. Most of the advancement in knowledge has taken place because of experiments conducted with the help of statistical methods. For example, experiments about crop yields and different types of fertilisers and different types of soils or the growth of animals under different diets and environments are frequently designed and analysed with the help of statistical methods. Statistical methods also affect research in medicine and public health. In fact, there is hardly any research work today that one can find complete without statistical data and statistical methods. Also it is impossible to understand the meaning and implications of most of the research findings in various disciplines of knowledge without having at least a speaking acquaintance with the subject to statistics.

(vii) Statistics and Other Uses

We have discussed above the significance of statistics in some important fields. Besides these, statistics are useful to various institutions such as bankers, brokers insurance companies, auditors, social workers, labour unions, trade associations and chambers of commerce. The banks have to make a very careful study of the cash requirements otherwise they may find they are short of cash and their existence is at stake. Similarly, the premium rates of the life insurance companies are based upon very careful study of the expectation of life.

Statistics are immensely useful to politicians and their supporters.They want to find out the prospects of winning the election and the efforts required for it. By sampling a certain percentage of voters prior to election, one can work out the percentage of votes the candidates is likely to receive in the election. The estimated percentage can then be used to decide, for example, whether a greater campaign is required to assure the candidate's victory.

These references to statistical applications are not intended to be exhaustive, but they simply suggest the diversity of applications of the underlying methods and ideas of statistics. In fact the applications of statistics are so numerous that statistics today has risen from the science of statecraft to the science of universal applicability. It is instrumental in enhancing human welfare and is such a master-key that enables to solve the problems of mankind almost in every field. Most of the people make use of statistics consciously or unconsciously in taking decisions. Statistical knowledge is in fact essential for a good citizen. H.G. Wells was right when he said '*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.*'

It must be remembered that the *statistical approach, though universal in its underlying ideas, must be tailored to fit the peculiarities of each concrete problem to which it is applied. It is dangerous to apply statistics in cookbook style, using the same recipes over and over, without careful study of the ingredients of each new problem.*

Also the reader must understand that statistics is not a dry, abstract and unrealistic pursuit followed by a small group of highly trained mathematicians, but rather a vitally important part of the economic and business life of the community. The usefulness of statistics to the reader depends to a great extent on his ability to use his imagination in applying the various statistical tools to his own particular situation.

STATISTICS AND THE COMPUTER

It may be interesting to note that the development of statistics has been closely related to the evolution of electronic computing machinery. Statistics is a form of data processing, a way of converting data into information useful for decision-making. Processing of 'raw data' is extensively required in the application of many statistical techniques. Statistical theory is generally expressed in the form of mathematical equations. However, the applications of this theory requires processing of real data.

As statistical theories become more complex, it becomes increasingly difficult to perform the calculations needed to apply these theories. Hence, in one sense the development of statistical theory and electronic computers reinforce each other. As statisticians devise new ways of describing and using data of decisions, computer scientists respond with newer more efficient ways of performing these operations. Conversely, with the evolution of more powerful computing techniques, people in statistics are encouraged to explore new and more sophisticated methods of statistical analysis. The computers can process large amounts of data quickly and accurately. This is a great benefit to businesses and other organisations that must maintain records of their operations. The computer brings efficient data processing to familiar operations such as payroll calculations, inventory management and airline reservation system. With the advancement in computer technology more and more people coming in direct contact with computers. New microprocessors have made the home computer a reality for both work and entertainment. It may be pointed out that the output from a computer is only as good as the data input ("Garbage in, garbage out" is the popular saying). This warning applies equally to statistical analysis. Statistical decisions based on data are no better than the data used.

LIMITATIONS OF STATISTICS

Despite the usefulness of statistics in many fields, impression should not be carried that statistics are like magical devices which always provide the correct solution of problems. Unless the data are properly collected and critically interpreted there is every likelihood of drawing wrong conclusions. Therefore, it is also necessary to know the limitations and the possible misuses of statistics. The following are the important limitations of the science of statistics :

1. *Statistics does not deal with isolated measurement.* Not all quantitative data are statistical. Isolated measurements are not statistical. Data are statistical when they relate to measurement of masses, not statistical when they relate to an individual item or event as a separate entity. For example, the wage earned by an individual worker at any one time, taken by itself, is not a statistical datum but taken as a part of a mass of information, it may be a statistical data. It should be noted that the wages of one worker over a period of time, being a series of wages, can be used statistically.

2. *Statistics deals only with quantitative characteristics.* Statistics are numerical statements of facts. Such characteristics as cannot be expressed in numbers are incapable of statistical analysis. Thus, qualitative characteristics like honesty, efficiency, intelligence, blindness and deafness cannot be studied directly. However, it may be possible to analyse such problems statistically by expressing them numerically. For example, we may study the intelligence of boys on the basis of the marks obtained by them in an examination.

3. *Statistical results are true only on an average.* The conclusions obtained statistically are not universally true; they are true only under certain conditions. This is because statistics as a science is less exact as compared to natural sciences.

4. *Statistics is only a means.* Statistical methods furnish only one method of studying a problem. They may not provide the best solution under all circumstances. Very often it may be necessary to supplement the conclusions arrived at by the help of statistics with the other methods that may be used to study a problem. It should be carefully noted that statistics is only a means and not an end. It analyses the facts and throws light on the real situation. In deciding a course of action it may be necessary to take into account the country's culture, religions, philosophy, personal, political or other non-quantitative considerations. Exclusive dependence on statistics may lead to fallacious conclusion in many situations.

5. *Statistics can be misused.* The greatest limitation of statistics is that it is liable to be misused. The misuse of statistics may arise because of several reasons. For example, if statistical conclusions are based on incomplete information, one may arrive at fallacious conclusions. Thus the arguments that drinking beer is bad for longevity since 99% of the persons who take beer die before the age of 100 years is statistically defective, since we are not told what percentage of persons who do not drink beer and die before reaching that age. Statistics are like clay and they can be moulded in any manner so as to establish right or wrong conclusion. Moreover, any Tom, Dick and Harry cannot deal with statistics. It requires experience and skill to draw sensible conclusions from the data ; otherwise, there is every likelihood of wrong interpretations. Also statistics cannot be used to full advantage in the absence of proper understanding of the subject to which it is applied.

Distrust of Statistics

By distrust of statistics we mean lack of confidence in statistical statements and statistical methods. It is often believed that "Statistics can prove anything." "There are three types of lies—lies, damn lies and statistics—wicked in the order of their naming." The following three main reasons account for such notions being held by people about statistics :

1. Figures are convincing and, therefore, people are easily led to believe them.

2. They can be manipulated in such a manner as to establish foregone conclusions.

3. Even if correct figures are used, these may be presented in such a manner that the reader is misled. For example, note the following statement : "The profits of firm *A* are Rs. 80 crore for the year 2009-10 and that of firm *B* Rs. 98 crore for the same period." On the basis of this information only one would form the opinion that firm *B* is better than firm *A*. However, if we examine the amount of capital invested in both the firms, the quality or work done, etc., we might reach a different conclusion.

It should be noted that statistics neither proves anything nor disproves anything. It is only a tool. If properly used, tools can do wonders and, if misused, can be disastrous. The same is true of statistical tools. If used properly, they help in taking wise decisions and if misused they can do more harm than good. But the fault does not lie with the science of statistics as such.

PROBLEMS

1-A. Answer the following questions, each question carries **one** mark :

(*i*) What is business statistics ?

(*ii*) Give any two uses of statistics.

(*iii*) Can statistics prove anything ?

(*iv*) Comment : Figures do not die but liars figure.

(*v*) Can single and isolated figures be called statistics ?

(*vi*) What are the limitations of business statistics ?

(*vii*) Why there is lot of distrust about statistics ?

(*viii*) Is statistics science or art ?

(*ix*) How statistics are useful to managers ?

(*x*) Is comparison of statistical data desirable ?

4. *Statistics is only a means.* Statistical methods furnish only one method of studying a problem. They may not provide the best solution under all circumstances. Very often it may be necessary to supplement the conclusions arrived at by the help of statistics with the other methods that may be used to study a problem. It should be carefully noted that statistics is only a means and not an end. It analyses the facts and throws light on the real situation. In deciding a course of action it may be necessary to take into account the country's culture, religions, philosophy, personal, political or other non-quantitative considerations. Exclusive dependence on statistics may lead to fallacious conclusion in many situations.

5. *Statistics can be misused.* The greatest limitation of statistics is that it is liable to be misused. The misuse of statistics may arise because of several reasons. For example, if statistical conclusions are based on incomplete information, one may arrive at fallacious conclusions. Thus the arguments that drinking beer is bad for longevity since 99% of the persons who take beer die before the age of 100 years is statistically defective, since we are not told what percentage of persons who do not drink beer and die before reaching that age. Statistics are like clay and they can be moulded in any manner so as to establish right or wrong conclusion. Moreover, any Tom, Dick and Harry cannot deal with statistics. It requires experience and skill to draw sensible conclusions from the data ; otherwise, there is every likelihood of wrong interpretations. Also statistics cannot be used to full advantage in the absence of proper understanding of the subject to which it is applied.

Distrust of Statistics

By distrust of statistics we mean lack of confidence in statistical statements and statistical methods. It is often believed that "Statistics can prove anything." "There are three types of lies—lies, damn lies and statistics—wicked in the order of their naming." The following three main reasons account for such notions being held by people about statistics :

1. Figures are convincing and, therefore, people are easily led to believe them.

2. They can be manipulated in such a manner as to establish foregone conclusions.

3. Even if correct figures are used, these may be presented in such a manner that the reader is misled. For example, note the following statement : "The profits of firm *A* are Rs. 80 crore for the year 2009-10 and that of firm *B* Rs. 98 crore for the same period." On the basis of this information only one would form the opinion that firm *B* is better than firm *A*. However, if we examine the amount of capital invested in both the firms, the quality or work done, etc., we might reach a different conclusion.

It should be noted that statistics neither proves anything nor disproves anything. It is only a tool. If properly used, tools can do wonders and, if misused, can be disastrous. The same is true of statistical tools. If used properly, they help in taking wise decisions and if misused they can do more harm than good. But the fault does not lie with the science of statistics as such.

PROBLEMS

1-A. Answer the following questions, each question carries **one** mark :

(*i*) What is business statistics ?

(*ii*) Give any two uses of statistics.

(*iii*) Can statistics prove anything ?

(*iv*) Comment : Figures do not die but liars figure.

(*v*) Can single and isolated figures be called statistics ?

(*vi*) What are the limitations of business statistics ?

(*vii*) Why there is lot of distrust about statistics ?

(*viii*) Is statistics science or art ?

(*ix*) How statistics are useful to managers ?

(*x*) Is comparison of statistical data desirable ?

1-B. Answer the following questions, each question carries **four** marks :

(*i*) Explain some important functions of statistics.

(*ii*) With the help of few examples point out the role of statistics in Business and management.

(*iii*) How statistics and computers are related ?

(*iv*) "Statistics is the foundation of sound decision-making". Elucidate giving suitable examples.

(*v*) What are the limitations of statistics ?

2. Define statistics. How does it help a manager ?

3. How far can statistics be applied for business and management decisions ? Discuss briefly bringing out limitations, if any. [*MBA, Delhi Univ., 2001*]

4. What is statistics ? How do you think the knowledge of statistics is essential in management decisions. Illustrate your answer through examples.

5. Are statistical methods likely to be of any use to a marketing firm ? Illustrate your answer with some typical marketing problems and the statistical techniques to be used there. [*MBA, Roorkee Univ., 2000, MBA, Delhi Univ., 2002, 2007*]

6. Comment on the following statements :

(*i*) "Figures do not lie but liars figure."

(*ii*) "The science of statistics is a most useful servant but only of great value to those who understand its proper use."

(*iii*) "Statistics is the science of averages." [*MBA, HPU, 2004*]

7. "Statistics is a body of method for making wise decisions in the face of uncertainty." Comment on the statement bringing out how clearly does statistics help in business decision-making. [*MBA, Osmania Univ., 2006*]

8. "There are three kinds of lies : lies, damn lies and statistics." Comment on this statement and point out the limitations of statistics.

9. (*a*) "Statistics is all-pervading." Elaborate.

(*b*) "Statistics is what statisticians do." Examine critically.

10. "Statistics are numerical statements of facts but all facts numerically stated are not statistics." Comment upon the statement.

11. How will you explain in brief the meaning of statistics to a layman ?

12. Define statistics, and statistical methods. Explain the uses of statistical methods in modern business organizations. [*MBA, Vikram Univ., 2005*]

13. Critically examine the following statements :

(*a*) "Statistics can prove anything."

(*b*) "Statistics only furnishes a tool, necessary though imperfect."

(*c*) Explain how statistics plays an important role in management planning and decision-making.

14. Discuss briefly the applications of business statistics pointing out their limitations, if any.

15. Describe the main areas of business and industry where statistics are extensively used. [*MBA, Delhi Univ., 2007*]

16. "Statistics are like clay of which you can make a God or Devil as you please." In the light of this statement discuss the uses and limitations of statistics.

17. With the help of a few examples explain the role of statistics as a managerial tool.

18. "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." Comment. Also give two examples of how the science of statistics could be of use in managerial decision-making. [*MBA, HPU., 2005*]

19. Whether the statements are *true* or *false* : (*i*) Statistics are affected to marked extent by a multiplicity of causes, (*ii*) No volume of statistics can replace the knowledge and experience of executives.

20. "Statistics is a method of decision-making in the face of uncertainty on the basis of numerical data and calculated risks." Comment and explain with suitable illustration.

21. "Statistical Methods are most dangerous tools in the hands of the inexperts." Examine this statement. How are statistics helpful in business and industry ? Explain.

22. (*a*) Define statistics. Discuss its applications in the management of business enterprises. What are its limitations, if any ? [*MBA, Jodhpur Univ.; MBA, HPU, 2007*]

(*b*) "Without adequate understanding of statistics, the investigator in social sciences may frequently be like the blind man groping in a dark closet for a black cat that is not there." Comment.

23. (*a*) Explain the utility of statistics as managerial tool. Also discuss its limitations.
    (*b*) "Modern statistical tools and techniques are basically important for improving the quality of managerial decisions." Explain this statements and discuss the role of statistics in planning and control of business. [*MBA , HPU, 2008*]
24. What role does Business Statistics play in the management of a business enterprise ? Examine its scope and limitations.
25. "The fundamental gospel of statistics is to push back the domain of ignorance, rule of thumb, arbitrary or premature decisions, traditions and dogmatism and to increase the domain in which decisions are made and principles are formulated on the basis of analysed quantitative facts." Explain the statement with the help of a few business examples.
    [*MBA , Osmania Univ., 2002*]
26. "Quantitative tools and techniques are basically important for improving the quality of managerial decisions." Examine the statement and discuss the role of quantitative techniques in planning and control of business activity.
    [*MBA , KU, 2009*]

*****

2

Collection of Data

Data constitute the foundation of statistical analysis and interpretation. Hence the first step in statistical work is to obtain data. Data can be obtained from three important sources, namely : (*i*) Secondary Source, (*ii*) Internal Records, and (*iii*) Primary Source. Depending on the source, we can have either secondary data or internal data or primary data.

(1) Secondary Data

Like all scientific pursuits, in statistics also the investigator need not begin from the very beginning, he may use and must take into account what has already been discovered by others. Consequently before starting a statistical investigation we must read the existing literature and learn what is already known of the general area in which our specific problem falls. When an investigator uses the data which has already been collected by others, such data are called *Secondary data.* Secondary data can be obtained from journals, reports, government publications, publications of research organisations, trade and professional bodies, etc. However, secondary data must be used with utmost care. The user should be extracautious in using secondary data and he should not accept it at its face value. The reason is that suchdata may be full of errors because of bias, inadequate size of the sample, substitution, errors of definition, arithmetical errors, etc. Even if there is no error, secondary data may not be suitable and adequate for the purpose of the inquiry. Hence, before using secondary data the investigator should examine the following aspects :

(*a*) *Whether the data are suitable for the purpose of investigation.* Before using secondary data the investigator must ensure that the data are suitable for the purpose of the inquiry. The suitability of data can be judged in the light of the nature and scope of investigation. For example, if the object of inquiry is to study the wage levels including allowances of workers and the data relate to basic wages alone, such data would not be suitable for the immediate purpose.

(*b*) *Whether the data are adequate for the purpose of investigation.* If it is found that the data are suitable for the purpose of investigation they should be tested for adequacy. Adequacy of the data is to be judged in the light of the requirements of the survey and the geographical area covered by the available data. For example, in the illustration given above, if our object is to study the wage rates of the workers in the sugar industry in India and if the available data cover only U.P., it would not serve the purpose. The question of adequacy may also be considered in the light of the time period for which the data are available. For example, for studying trend of prices we may need data for the last 8-10 years but from the sources known to us may be available for 5-6 years only which would not serve our object.

(*c*) *Whether the data are reliable.* To determine the reliability of secondary data is perhaps the most important and at the same time most difficult job. The following tests, if applied, may be helpful to determine how far the given data are reliable :

(*i*) Was the collecting agency unbiased or did it "have an axe to grind" ?

(*ii*) If the enumeration was based on a sample, was the sample representative ?

(*iii*) Were the enumerators capable and properly trained ? Incompetent or poorly trained enumerators cannot be depended upon to produce useful results.

(*iv*) Was there a proper check on the accuracy of field work ?

(*v*) Was the editing, tabulating and analysis carefully and conscientiously done ? Carelessness in either one or more of these functions can render of little value the findings of an otherwise valuable study.

(*vi*) What degree of accuracy was desired by the compiler ? How far was it achieved ?

(2) Internal Data

Internal data refer to the measurements that are the by-product of routine business record keeping like accounting, finance, production, personnel, quality control, sales, R & D, etc.

In statistical analysis of many business problems one may be able to use internal data which emerges in the process of keeping records such as employee earnings from a payroll, sales amounts from a sales journal, the amount of raw materials, direct labour and manufacturing expenses used and the units of finished product produced from production records, and cash receipts from the cash book. Thus the chief source of internal data are the accounting records kept in most business firms.

Since internal data originate within the business, collecting the desired information does not usually offer much difficulty. The particular procedure depends largely upon the nature of facts being collected and the form in which they exist. The problem of collection is primarily that of having the proper record made at the time the information is secured. The information wanted is frequently to be found in more than one department of the business, which increases the difficulty of getting just the information one wants.

(3) Primary Data

Primary data are measurements observed and recorded as part of an original study. When the data required for a particular study can be found neither in the internal records of the enterprise, nor in published sources, it may become necessary to collect original data, *i.e.*, to conduct first hand investigation. The work of collecting original data is usually limited by time, money and manpower available for the study. When the data to be collected are very large in volume, it is possible to draw reasonably accurate conclusions from the study of a small portion of the group called a *sample*. The actual procedures used in collecting data are essentially the same whether all the items are to be included or only some items are considered.

There are two basic methods of obtaining primary data, namely :

(1) Questioning, and

(2) Observation.

Questioning, as the name suggests, is distinguished by the fact that data are collected by asking questions from people who are thought to have the desired information. Questions may be asked in person, or in writing. A formal list of such questions is called a *questionnaire**.

---

* A distinction is often made between a questionnaire and a schedule. Questionnaire refers to a device for securing answers to questions by using a form which the respondent fills in himself. Schedule is the name usually applied to a set a questions which are asked and filled in a face-to-face situation with another persons.

When data are collected by observation, the investigator asks no questions. Instead, he observes the objects or actions or actions in which he is interested. Sometimes individuals make the observations ; on other occasions, mechanical devices observe and record the desired information.

Observation method does not automatically produce accurate data. Physical difficulties in the observation situation on the part of the observer may result in errors. Even more important, however, is the influence on observations of the observer's training, philosophy, opinions and expectations. This is borne by the fact that significant variations in observation of the same phenomena have been reported for such diverse projects as the reading of X-ray films, E.C.G., state of repair for roads, etc.

Of the two methods named above, the questionnaire method is more widely used for collecting business data. When questionnaire method is used, three different techniques of communication with questionnaires are available : (1) personal interview, (2) mail, and (3) telephone. Personal interviews are those in which an interviewer obtains information from respondents in face-to-face meetings. The information obtained by this method is likely to be more accurate because the interviewer can clear up doubts, can cross-examine the informants and thereby obtain correct information. In most mail surveys, questionnaires are mailed to the respondents who are supposed to fill them and return by post. Sometimes, however, mail questionnaires are placed in respondent's hands by other means such as attaching them to consumer products, putting them in magazines or newspapers or having field workers leave them with respondents. In each case respondents complete the questionnaires themselves and send back the completed forms by post. This method has a special advantage in surveys where field of investigation is very vast and the informants are spread over a wide geographical area. Telephone interviews are similar to personal interviews except that communication between interviewer and respondent is on the telephone instead of direct personal contact. However, this method has several limitations such as it cannot be used to interview those people who don't have telephone, telephone conversation cannot be very long and also replies on the telephone can be very long and also replies on the telephone can be very erratic and unreliable.

The greatest advantage of the questionnaire method is the versatility as many of the business problems can be dealt with without much difficulty. Also, questioning is usually faster and cheaper than observing. The most important limitation of the questionnaire method is the difficulty in obtaining information from the respondents. The interviewer is unknown to the respondent, and the subject of the proposed interview may be of little or no interest. Questions about income or very personal subjects frequently meet refusals. Besides the difficulty of obtaining information, the respondents may supply wrong information. Also it is difficult, if not impossible, to state a given question in such a way that it will mean exactly the same thing to every respondent. Much of the success of the questionnaire method depends on the wisdom with which the questionnaire has been drafted. The designing of questionnaire involves many vital issues and is discussed in detail below.

DESIGNING A QUESTIONNAIRE

The success of the questionnaire method of collecting information depends largely on the proper designing of the questionnaire. Designing questionnaire is a highly specialised job and requires a great deal of skill and experience. It is difficult tolay down any hard and fast rules to be followed in this connection. Although much progress has been made, the designing of questionnaires is still very much an art.

Most of what is known about making questionnaires is based on experience. Neither a basic theory nor even a fully systematised approach to the problem has been developed. Nevertheless, the extensive experience of many researchers and a limited number of organised experiments have led to a considerable understanding of the problem and to a long list of "do's and don'ts" rules of thumb. These can definitely help a beginning researchers avoid pitfalls, but they cannot be substituted for creative imagination in designing a questionnaire procedure.

While developing a questionnaire, the researcher has to be very clear on the following issues :

1. What information will be sought ?
2. What type(s) of questionnaire will be required ?
3. How that (those) questionnaire (s) will be administered ?
4. What the content of the individual question will be ?
5. What the form of response of each question will be ?
6. How many questions will be used and how the individual questions will be sequenced ?
7. Whether the questionnaire shall be disguised or undisguised ?
8. Whether the questionnaire shall be structured or unstructured ?

STRUCTURED AND UNSTRUCTURED QUESTIONNAIRES

A questionnaire can be either structured or unstructured and disguised or undisguised as can be seen from the following :

	Structured	*Unstructured*
Undisguised	*A*	*B*
Disguised	*D*	*C*

Structure refers to the degree of standardisation imposed on the questionnaire. A higly structured questionnaire is one in which the questions to be asked and the response permitted are completely predetermined. A highly unstructured questionnaire is one in which the questions to be asked are only loosely predetermined, and the respondent is free to respond in his/her own words and in any way he/she sees fit. Unstructured techniques have two major disadvantages :

(1) They are slow and, hence, costly to administer in the field and to tabulate ; and

(2) The data collection process and the interpretation of results are both subjective and, hence open to bias. Structured techniques overcome these problems, but they are difficult to use in situations where respondents may hesitate to report their attitudes.

A disguised questionnaire attempts to hide the purpose of the study whereas an undisguised questionnaire is one in which the purpose of the research is obvious from the questions posed.

Structured undisguised questionnaires are the most commonly used type in practice. In such questionnaires the responses as well as the questions are standardised. This is accomplished by employing fixed alternative questions in which the responses of the subject are limited to the stated alternatives. An example of this type of questions regarding people's attitude towards social security and the more government legislation controlling it could be :

Do you feel India needs more (or less) social security legislation ?

- ☐ Needs more
- ☐ Needs less
- ☐ Neither more or less
- ☐ No opinion.

Structured undisguised questionnaires are simple to administer and easy to tabulate and analyse. The respondent also feels almost no difficulty in replying. The question "What is your marital status" is more confusing than is the question. "Are you married, single widowed or divorced ?" The fixed alternative questions are most productive when the possible replies are well known, limited in number and clear cut.

The unstructured undisguised questionnaire is one in which the purpose of the study is not concealed but the response to the question is open-ended. Thus consider the question "How do you feel about the need for legislation for more social security measures ?" Such questions provide complete freedom to the respondent. However, the responses are difficult to tabulate and analyse.

In the unstructured disguised questionnaires, the respondents are not directly told about the purpose of study and the questions are framed in a manner that there is complete freedom for the respondent to answer. The basic philosophy underlying such questionnaires is that the more unstructured and ambiguous a stimulus, the more a subject can and will project his emotions, needs, motivations, attitudes and values. The practical difficulties of editing, coding and tabulation of replies impose serious limitations on the use of the methods. This method is more often used for exploratory research than for descriptive or casual research.

The structured disguised questionnaires are also not very popularly used in practice. They emerged as an attempt to secure the advantages of disguise in revealing unconscious and hidden motives and attitudes along with the advantages in coding and tabulation common to structured questionnaires. The main advantages of this approach emerges in analysis.

Having decided these issues, the following points may be kept in mind while designing a questionnaire :

1. *Covering letter.* The person conducting the survey must introduce himself and state the objective of the survey. It is desirable that—

(*i*) A short letter is enclosed. The letter should state in as few a words as possible, the purpose of the survey and how the informant would tend to benefit from it.

(*ii*) Enclose a self-addressed envelope for the respondent's convenience in returning the questionnaire.

(*iii*) Assure the respondent that his answers will be kept in strictest confidence.

(*iv*) Promise the respondent that he will not be harassed after he fills up the questionnaire.

(*v*) Offer special inducement (free gifts, concession coupon, etc.) to return the questionnaire.

(*vi*) If the respondent is interested, promise him a copy of the result of the survey.

2. *The number of questions should be as few as possible.* The number of questions should be kept to the minimum. The precise number of questions to be included would naturally depend on the object and scope of the investigation. Once the objectives are clearly defined only question pertinent to the objectives should be asked. The time of the respondent should not be wasted by asking irrelevant questions. Fifteen to twenty-five may be regarded as a fair number. If a lengthy questionnaire is unavoidable, it should preferably be divided into two or more parts.

3. *Questions should be logically arranged.* The questions must be arranged in a logical order so that a natural and spontaneous reply to each is induced. They should not skip back and forth from one topic to another. Thus it is undesirable to ask a man how many children he has before asking whether he is married or not. Similarly, it would be illogical to ask a man his income before asking him whether he is employed or not. Thus, the sequence of the questions should be considered carefully in terms of the purpose of the study and the persons who will supply the information. Questions supplying identification and description of the respondent should come first followed by major information questions. If opinions are requested, such questions should usually be placed at the end of the list. Two different questions may be included on the same subject to provide a cross-check on important points.

4. *Questions should be short and simple.* The questions should be short, simple and easy to understand and they should convey one and only one sense. Unless the persons being interrogated is technically trained, technical terms should be avoided. Words such as "Capital" or "income" that have different meanings for different persons should not be used unless a clarification is included in the questions.

5. *Questions of a sensitive nature should be avoided.* As far as possible questions of a personal and pecuniary nature should not be asked. For example, questions about sources of income, volume of sales, etc., may be willingly answered in writing. Where such information is essential, it should be obtained indirectly, preferably personal interviews. For example, we may ask the respondent to indicate his salary, profits or sales turnover from among a set of ranges. Even then, such questions should be asked at the end of the interview when the informants feel more at ease with the interviewer.

6. *Instructions to the informants.* The questionnaire should provide necessary instructions to the informants. For example, the questionnaire should specify the time within which it should be sent back and the place where it should also be given. For instance, if there is a question on weight it should be specified as to whether weight is to be expressed in pounds or kilograms.

7. *Footnotes.* If a particular questions needs clarification, it should be marked or lettered and the explanation provided in footnotes.

8. *Questions should be capable of objective answers.* Various types of questions that may form part of a questionnaire can be grouped under three categories :

(*a*) dichotomous questions,

(*b*) multiple choice questions, and

(*c*) open-ended or free answer questions.

Dichotomous questions are fixed alternative questions in which only two alternatives are listed. The respondent has to tick one of these alternatives. Such questions can usually be answered in 'yes' or 'no'. Two examples of dichotomous questions are :

Do you intend to purchase a coloured television set this year ?

☐ Yes ☐ No

Are you satisfied with the after-sales service provided by our organisation ?

☐ Yes ☐ No

This is an excellent techniques if applied to situations where a clear cut alternative exists. For example, a question : 'Do you have a television ?' can be easily answered in 'yes' or 'no' questions be avoided or additional answers such as sometimes, cannot say, etc., must be included. For example, in order to find out which particular toothpaste people use, giving only two alternatives 'yes' or 'no' will not be enough because there may be some persons who are using it occasionally. This questions should be framed as follows :

(*a*) Do you use Colgate toothpaste ?

☐ Yes ☐ No

(*b*) If yes, how often

always .....; occasionally ;

seldom .....; never .....

In the multiple choice questions, the respondent is asked to select one out of a number of alternative responses. All possible answers to a question are listed and the respondent choose one of these. This process not only facilitates tabulation of data, but also takes very little time to the respondent to fill in the questionnaire. Thus, while ascertaining how do MBA students normally travel to Faculty of Management Studies, instead of asking 'how do you normally travel to the Faculty', frame a question of this type :

How do you normally travel to Faculty of Manangement ? (Tick)

(*i*) By Bus ☐

(*ii*) By your own car ☐

(*iii*) By your own scooter ☐
(*iv*) By taxi ☐
(*v*) By three wheeler scooter ☐
(*vi*) On foot ☐
(*vii*) Any other. ☐

Another example of a multiple choice questions is :

Why did you purchase Onida TV ?

☐ Price is lower than other brands.
☐ It represents best quality.
☐ Picture is better.
☐ Warranty period is longer.
☐ After-sales service is better.
☐ Any other.

The problem with such questions is that the respondent may like to tick more than one alternative. For example, one might have bought that brand of TV not only because of lowest prices, but longer period of warranty and best after-sales service. Hence, the respondent should be instructed to 'check the most important reason', 'check all those reasons that apply' or 'rank all the reasons that apply from most important to least important'. However, this type of question is excellent if most of the possible answers are both known and few in number. When the possible answers are numerous, a limited list—even accompanied by 'any other' category—may elicit a response different from that which otherwise would be forthcoming. The use of multiple choice questions is indicated only when the investigator is confident of the existence of a limited group of important alternatives and it should be avoided when there are many possible responses of relative equal significance.

In the free answer form or open-ended questions, the respnses is asked to answer a question in his/her own words in essay form. The MBA students after completion of the course may be asked questions like :

What is your opinion of the quality of teaching ?

What do you feel about the facilities offered by the faculty ?

What do you think of the practical usefulness of the course you have undergone ?

The difficulty with the free answer questions is in classifying the responses. This is often difficult, time-consuming and somewhat arbitrary.

In most questionnaires one may find it necessary to employ all the three types of questions to elicit the information required.

9. *Answer to questions should not require calculations.* Questions should not require calculations to be made. For example, informant should not be asked yearly income, for in most cases they are paid monthly. Similarly, questions necessitating calculation of ratio and percentages, etc., should not be asked as it may take much time and the informant may not send back the questionnaire at all.

10. *Pre-testing the questionnaire.* The questionnaire should be pre-tested with a group before mailing it out. The advantage of pre-testing is that the shortcomings of the questionnaire can be discovered and it can be revised in the light of the try out.

11. *Cross-checks.* If possible, one or more cross-checks should be incorporated into the questionnaire, to determine whether the respondent is answering the questions carefully.

12. *Incentives to respondents.* Some incentive for filling up the questionnaire should be provided. It may be in the form of gift coupons, a sample of a product which the company wanted to introduce, etc. Sometimes even a promise to supply a copy of the findings after the survey work is over, works as an incentive.

13. *Method of tabulation to be used.* The method to be used for tabulating the results should be determined before the final draft of the questionnaire is made. These days most of the surveys are conducted on a large scale. This necessitates the use of computers. When the results are to be computerized, the questionnaire has to be drafted in a different way. This does not mean that the basic principles of constructing questionnaire are changed—the only change is that every question is to be properly coded. It is suggested that whenever the results are to be processed on computer, the guidance or the computer expert should be obtained before the questionnaire is finally printed.

It should be kept in mind that though the above points provide a guide or a checklist researchers can follow in their first encounters with the problem of questionnaire design, blind adherence to the above procedure shall do more harm than good. With questionnaires, the "proof of pudding lies very much in the eating," *i.e.,* a questionnaire shall serve its purposes if it is able to produce accurate data of the kind needed. The proper construction of a questionnaire is a skill which is generally developed only by experiences in the use of research methodology or by on the job training. The natural tendency to rush through the construction should be avoided. Time spent in this stage of a well planned survey or experiment is invariably found to be extremely valuable in retrospect.

PRE-TESTING THE QUESTIONNAIRE

Pre-testing the questionnaire occupies a place of great significance in a survey. A researcher should not expect that the first draft of his/her efforts will result in a usable questionnaire. The researchers should examine each question with a jaundiced eye to assure that the question is not confusing or ambiguous, potentially offensive to the respondent, leading to biased responses, etc. The real test of a questionnaire is how it performs under actual conditions of data collections. For this assessment, the questionnaire pre-test is vital. The questionnaire pre-test serves the same role in questionnaire design as test marketing serves in new product development. Test marketing provides the real test of customer reactions to the product and the accompanying marketing programme. Similarly, the pre-test provides the real test of the questionnaire and the mode of administration. Some of the advantages of pre-testing the questionnaire are :

1. The investigator can find out what are the shortcomings of the questionnaire. Even the best designed questionnaire may have some problems. For example, there may be ambiguous questions, sequence may require changes, some questions may have to be dropped, some questions may have to be asked in different froms and still some new questions may have to be added. The time to know about all these problems is before the full-scale survey or experiment is conducted—not after.

2. An idea can be formed about the extent of non-response likely to take place.

3. Greater co-operations of the informants can be secured. Even persons most allergic to writing can with proper inducement be persuaded to answer the questionnaire. It is the surveyor's job to find out what these appeals are.

While pre-testing the questionnaire, it is desirable to cover a cross-section of the population eventually to be surveyed. When the sample is drawn, it should be broken down into various sub-samples by taking, for instance, every tenth or every hundredth case from the entire list.

The pre-test should be done by the personal interview regardless of the actual mode of administration that will be used. The work of pre-testing must be done with utmost care and caution otherwise unnecessary

and unwanted changes may be introduced. The firm's most experienced interviewers should be employed to conduct the pre-test. If the pre-test suggests major changes in the question, the questionnaire should again be pre-tested employing personal interviews. If the changes are minor, the questionnaire can then be pre-tested second time using mail or telephone.

After each significant revision of the questionnaire, another pre-test should be run. When the last pre-test suggests no new revisions, the researcher should get the questionnaire finally printed.

It is desirable that the response resulting from pre-test be coded and tabulated. The tabulation of pre-test responses can serve as a check on our conceptulisation of the problem and the data and method of analysis necessary to answer it. It is said that if a researcher who avoids a questionnaire pre-test and tabulation of replies is either a naive or a fool. The pre-test is the most inexpensive insurance the researcher can buy to ensure the success of the questionnaire and the research project.

SPECIMEN QUESTIONNAIRE

Two specimen questionnaires are given below. In the first, manual tabulation of results would be done and in the second, computerised results would be obtained.

A study* entitled 'Consumer Survey on Television Sets' was conducted by a post-graduate student of the Faculty of Management Studies. The basic objectives of carrying out this survey were to determine :

- The people who influence the purchase decision of a particular brand of television set.
- The people in the family who decide about the budget for the purchase of television set.
- The factors that influences the selection of a particular dealer/ showroom for the purchase.
- The various attributes of the production, which, if introduced in a particular brand, would create more market for the brand.
- Effectiveness of the I.S.I. mark in the context of purchase of a particular brand of TV set.
- The factors that influence the selection of a particular size of TV set.
- The economic profile of people who own TV sets.
- The importance of the TV set as a status symbol to an owner.
- The various factors that can be emphasized in providing the sales of a particular brand.
- The comparative effectiveness of promotion channels for TV sets.

The questionnaire designed for attaining the above objectives is given below :

QUESTIONNAIRE

1. Do you have Television Set ? Yes ☐ No ☐
   If yes,
   What brand ? .....
   What is the size of the screen ? .....
   When did you buy it ? (Month and Year) .....
   Approximate, price paid ? .....
2. Of the following persons, who advised you to purchase the above brand of TV set [tick the appropriate box(es)]:
   ☐ Yourself ☐ Your friend ☐ Your wife/husband ☐ Your neighbour
   ☐ Your children ☐ Your colleague ☐ Your parents ☐ Dealer
   ☐ Any other (Please state who)

* The study was conducted by Mr. Deepak Mahendru.

12. Who decided about the budget for TV set's purchase ? [tick the appropriate box(es)]

☐ Yourself ☐ Your parents ☐ Your wife/ husband ☐ Any other (please state who)

13. Who decided about the place you bought your TV set from ?

☐ Yourself ☐ Your friend ☐ Your wife/husband ☐ Your neighbour

☐ Your children ☐ Your colleague ☐ Your parents ☐ Any other (Please state who)

14. Do you remember the place from where you bought your TV set ?

☐ Yes ☐ No

If yes, please state the name and address of showroom or company (if bought direct from the company).....

15. Is the showroom near the house ? Yes ☐ No ☐

16. How did you come to know the above showroom/dealer/company ? [tick the appropriate box(es)] :

☐ Through radio advertisement ☐ Through newspaper ☐ Through TV advertisement

☐ Through cinema slides ☐ Through roadside hoardings ☐ Through hand bills

☐ Through magazines ☐ Through some person

17. Rank the following factors affecting decision to buy TV according to their importance ('1' of highest priority factor, '2' next important factor and so on up to '11') :

☐ Size of showroom ☐ Discount offered by dealer

☐ Market where showroom is located ☐ Free transformer and antenna

☐ Distance of showroom from house ☐ Free installation offered by dealer

☐ Reputation of showroom ☐ Rate of interest charged by dealer on hire-purchase instalments

☐ After-sales service offered by dealer

☐ Range of products sold in the showroom

☐ Gift at time of purchase offered by dealer.

18. Did you make a special survey of the market before buying your TV set ?

☐ Yes ☐ No

If yes, please state why (tick whichever boxes are appropriate) :

To see performance of various brands of TV sets ☐

To check which dealer offered lower price ☐

To check which dealer offered more incentives ☐

To check which dealer offered better after-sales service ☐

19. Do you get your TV set repaired by the same dealer from where you bought your TV set ?

☐ Yes ☐ No

20. In your opinion, which of the following media is giving the most effective and impact-creating advertisement (tick only two boxes) :

☐ Radio ☐ Magazines ☐ Television ☐ Newspapers

☐ Cinema slides ☐ Hoardings on roadside ☐ Advertisement films ☐ Hand bills

21. Where do you come across most of the advertisements concerning TV sets (tick only two boxes) ?

☐ Radio ☐ Magazines ☐ Television ☐ Newspapers

☐ Cinema slides ☐ Hoardings on road side ☐ Advertisement films ☐ Hand bills

22. How much time *per day*, on an average, do you spend in listening/reading the following ?

☐ Radio ☐ Newspapers ☐ Magazines

23. State the magazines/newspapers/periodicals you buy for your family.

24. Rank the following factors according to your preference at the time of purchasing the TV set (rank from 1 to 8) :

☐ Picture quality ☐ Shape of cabinet ☐ Price ☐ Material of cabinet

☐ Sound ☐ Colour of cabinet ☐ Size of screen ☐ Shutter/door

16. Would you prefer to put extra screen on your T.V. ?
☐ Yes ☐ No

If yes, it is because
(*i*) the TV screen without extra screen is harmful to eyes.
(*ii*) the double screen looks more elegant.
(*iii*) any other reason.

17. Which one do you think is better (tick *only one* box) ?
☐ Removable legs ☐ TV set in showcase ☐ Separate stand ☐ TV set on the table
Specify the reason for your choice from the following :
☐ Price advantage ☐ Ease of transportation ☐ Appearance ☐ Flexibility of placing TV set anywhere

18. Which, in your opinion, is better ?
☐ TV supported on legs ☐ TV supported on one single leg in the centre
State reasons for your choice (tick *only one* box) :
☐ Appearance ☐ Stability

19. Which one do you think is better ?
☐ Plastic moulded cabinet ☐ Wooden cabinet
State reason for your choice (tick *only one* box) :
☐ Low price ☐ Appearance ☐ Sturdiness

20. What shade of cabinet do you prefer ?
Wooden : ☐ Light ☐ Medium ☐ Dark
Plastic : ☐ Black ☐ Red ☐ Yellow
☐ White ☐ Blue ☐ Any other (specify)

21. Which type of 'Off-on' knob is better ?
☐ Rotating type ☐ Push-pull type

22. Which type of control knobs are better ?
☐ Rotating type ☐ Sliding type

23. Which type of channel selector appeares better ?
☐ Rotating type ☐ Push button type

24. Which would you prefer ?
☐ Antenna on roof top ☐ Indoor antenna

25. Do you think it better to have the voltage regulator within the TV set cabinet ?
☐ Yes ☐ No

26. Do you prefer a stand that can be used to adjust the height and direction of TV screen ?
☐ Yes ☐ No

27. Rank the following factors as per their importance to you at the time of your deciding the size of screen when you ar purchasing TV set ?
☐ Strain on eyes ☐ Size of room ☐ Size of picture on screen ☐ Appearance of TV set

28. Which one of the following types of TV sets is better in your opinion ?
☐ Valve set ☐ Semi-solid set ☐ IC set ☐ Don't know
State reason for your choice :
☐ Price advantage ☐ Low maintenance ☐ Lower power consumption ☐ Don't know

29. In your opinion, ISI mark is
☐ Symbol of quality ☐ Gimmick to attract customers ☐ Don't know

30. In which room have you kept your set ?
☐ Drawing room ☐ Bed room ☐ Dining room ☐ Study room
Please give approximate size of the room .....

III. Which one do you think is the best seller (rank ten of the following) ?

☐ Konark	☐ Bigston	☐ Uptron	☐ Standard
☐ Niki-Tasha	☐ Crown	☐ JK	☐ Bush
☐ Sylvania	☐ Televista	☐ Dyanora	☐ Disco
☐ Texla	☐ BPL	☐ Gem	☐ Weston
☐ Beltek	☐ Hotline	☐ Onida	

IV. Which of the following did you have *before* buying your TV ?

☐ Scooter	☐ Record player	☐ Transistor radio	☐ Car
☐ Cooking range	☐ Cassette player	☐ Refrigerator	☐ Radio
☐ Stereo record player	☐ Washing machine		

QUESTIONNAIRE 2*

(A study of problems of EXECUTIVE SELECTION and RETENTION in different organisations)

PARTICULARS OF EXECUTIVE/PROFESSIONAL RESPONDENTS

1. **NAME :** (Please fill only one letter in each box. Leave/ one box vacant between first and second name and second/ third/ surname and so on.)

2. **PHYSICAL FEATURES :**

(*i*) *AGE* in years (please tick mark (✓) the appropriate bracket)

Below 20	( )	20–25	( )	26–30	( )
31–35	( )	36–40	( )	41–45	( )
45–50	( )	51–55	( )	56–60	( )
Above 60					

(*ii*) *Height* ..... *cm.* *Weight* ..... *kg.*

(*iii*) *With Glasses* ( ) *Without Glasses* ( )

3. **SEX :** Male ( ) *Female* ( )

4. **EDUCATION :**

(*i*) Degree and Diplomas :

Bachelor's Degree	( )	Master's Degree	( )		
Diploma	( )	Graduate Diploma	( )	Post-Gr. Diploma	( )
Ph.D.	( )	Post Doctorate	( )		

Any other qualifications .....

Complete qualifications .....

(M.A. ; L.L.B. ; B.Sc. (Civil Engg.) ; M.B.A., etc.

(*ii*) Division :

Graduate	1st ( )	2nd ( )	3rd ( )
Post-graduate	1st ( )	2nd ( )	3rd ( )
Diploma	1st ( )	2nd ( )	3rd ( )
P.G. Diploma	1st ( )	2nd ( )	3rd ( )
.....	1st ( )	2nd ( )	3rd ( )
.....	1st ( )	2nd ( )	3rd ( )
.....	1st ( )	2nd ( )	3rd ( )

* This questionnaire was prepared by Shri S. Shamin Shah, a research scholar, Faculty of Management Studies, University of Delhi, in connection with his Ph.D. work. This in only one part of the questionnaire.

(*iii*) Subjects studied :

1. At Graduate level .....
2. At post-graduate level.....
3. At Ph.D. level.....
4. Others.....

(*iv*) Area of Specialisation :

1. At Bachelor's Degree level.....
2. At Master's Degree level.....
3. At Diploma/Post-graduate level.....
4. At Ph.D. level.....
5. At Post Doctorate level.....

5. EXPERIENCE : (Please indicate total relevant experience in years)

0—3 ( ) 4—7 ( ) 8—11 ( ) 12—15 ( ) Over 15 ( ).

Exact No. of years.....

6. EXPERIENCE ON PRESENT JOB :

0—3 ( ) 4—7 ( ) 8—11 ( ) 12—15 ( ) Over 15 ( )

Exact No. of years.....

7. PRESENT DESIGNATION :

.....

8. PRESENT EXECUTIVE/PROFESSIONAL LEVEL :

JUNIOR ( ) MIDDLE ( ) SENIOR ( ) TOP LEVEL ( )

Reporting to.....

9. PAY SCALE (If any) :

.....

EMOLUMENTS

10. BASIC PAY (per month) :

Below Rs. 700	( )	701—1100	( )	1101—1500	( )	1501—2000	( )
2001—2500	( )	2501—3000	( )	3001—3500	( )	3501—4000	( )
4001—4500	( )	4501—5000	( )	5001—5500	( )	5501—6000	( )
6001—6500	( )	6501—7000	( )	Above—7000	( )		

11. GROSS TOTAL PAY : (including Basic, DA, ADA, CCA, HRA, Conveyance and other Allowances)

1000—1500	( )	1501—2000	( )	2001—2500	( )	2501—3000	( )
3001—3500	( )	3501—4000	( )	4001—4500	( )	4501—5000	( )
5001—5500	( )	5501—6000	( )	6001—6500	( )	6501—7000	( )
7001—7500	( )	7501—8000	( )	Above 10,000	( )		

ALLOWANCES

12. Does your salary include following in addition to basic pay :

DA ( ) ADA ( ) CCA ( ) HRA ( ) CONVEYANCE ( )

ALLOWANCES

SPECIAL ALLOWANCES

13. If you get any other allowances/monthly monetary benefits, please mention names (not amounts) :

1. .....
2. .....
3. .....
4. .....

14. PERQUISITES AND OTHER FRINGE BENEFITS availed of by you :

(Please tick mark in the brackets applicable)

Free furnished flat/house	( )	Subsidised/leased or company accommodation	( )	Office Tele.	( )
				Direct Phone	( )
				PBX/PABX Extn.	( )

Residential telephone ( )	Domestic help	Cook ( )	Free Transport		
		Mail ( )	Bus ( )		
		Driver ( )	Car ( )		
			Scooter ( )		
House building advance ( )	Retirement benefits ( )		L.T.C. ( )		
Medical Free ( ) Treatment Subsidised ( )	Group term/ Executive Insurance Scheme ( )		Pension and post-retirement benefits ( )		
Provident Fund ( )	Free Education of Children	( )	Free conveyance for school going children		( )
Wife/Husband/ Sons job ( )	Leave encashment ( )				

OTHER PERKS

25. Please list other perquisites, benefits and facilities provided to you by your employers in additions to the above :

1. .....
2. .....
3. .....
4. .....
5. .....

26. Please indicate approximate value in rupees of these *other perquisites* (14 & 15) received in cash or kind per month over and above salary :

Below Rs. 1000	( )	1001—1500	( )	1501—2000	( )
2001—2500	( )	2501—3000	( )	3001—3500	( )
3501—4000	( )	4001—4500	( )	4501—5000	( )
5001—6000	( )	6001—7000	( )	7001—8000	( )
8001—9000	( )	9001—10000	( )	10001 & Above	( )

Please mention exact amount.....

27. **MARITAL STATUS :**

Single ( ) Married ( ) Any other specify ( )

28. Wife/Husband's Education :

Non-matric	( )	Matric	( )	Inter (10+2)	( )
Graduate	( )	Post-graduate	( )	Ph. D.	( )

Mention exact qualifications (like M.A., M.Ed., etc.).....

29. Her/his any other qualifications or training than above :

1. .....
2. .....
3. .....

30. Is your wife/husband also employed/self-earning ?

Yes ( ) No ( )

31. What is her/his profession/designation ?

Profession..... Designation.....

32. If employed, full time or part time ?

Full time ( ) Part time ( )

33. Please tick mark (√) her/his total monthly salary (Basic+DA+other allowances) or income :

Up to Rs. 500	( )	501—1000	( )
1001—1500	( )	1501—2000	( )
2001—2500	( )	2501—3000	( )
3001—3500	( )	3501—4000	( )
4001—5000	( )	Above 5000	( )

24. **COUNTRY OF DOMICILE :**
(1) Your's..... (2) Your spouse's.....

25. **MOTHER TONGUE :**
(1) Your's..... (2) Your spouse's.....

26. **LANGUAGES KNOWN** (Working knowledge) :
.....

27. **FAMILY :**
Single ( ) Husband and wife ( ) Husband, wife & children ( )

28. **CHILDREN :**
Sons :
One ( ) Two ( ) Three ( ) Four ( ) More ( )
Daughters :
One ( ) Two ( ) Three ( ) Four ( ) More ( )

29. (*i*) **DEPENDENTS :**
None ( ) Spouse and children ( ) Spouse.....children and parents/others ( )
(*ii*) **PARTIALLY DEPENDENT :**
One ( ) Two ( ) Three ( ) Four ( ) More ( )

30. **WHAT KIND OF FAMILY YOU HAVE ?**
Nuclear family (H.W. & C.) ( ) Joint/Extended family ( )

Background

31. **ORIGINALLY FROM :**
Village ( ) City ( )
Province.....

Job changes/trunover

32. How many jobs have you changed ?
One ( ) Two ( ) Three ( ) Four ( ) More ( )
Exact No. if more than four.....

33. Are you :
(*i*) going to change your present job ?
Yes ( ) No ( ) Don't know ( )
(*ii*) going to settle down now with this job ?
Yes ( ) No ( ) Don't' know ( )

34. Do you think :
(*i*) you were made for this job
Yes ( ) No ( ) Don't know ( )
(*ii*) you are doing this job because you could not get a better one
Yes ( ) No ( ) Don't know ( )
(*iii*) this is a good jumping board
Yes ( ) No ( ) Don't know ( )

JOB ABROAD

35. If given choice where would you like to work :
In India ( ) Abroad ( )
If abroad, mention 3 countries in order of preference :
1. ..... 2. ..... 3. .....

36. What are the 3 major considerations in your preferring a job abroad-select from the following :
More money ( ) Professional growth ( ) Job satisfaction ( ) Prestige ( ) Family circumstances ( ) Further education and training etc. ( ) Any other considerations.
1. ..... 2. ..... 3. .....
Not applicable ( )

37. If you go abroad for a job would you like to :

(*i*) Return home after working for 2-3 yrs. ( ) 5 yrs. ( )
10 yrs. ( ) More ( )

(*ii*) Settle down there permanently Yes ( ) No ( )
Not applicable ( ) Don't know ( )

38. What would you like to do after you return from a foreign assignment ?

Will try to get a job ( )
Will start own business ( )
Would relax ( )
Don't know ( )
Not applicable ( )

ABOUT PRESENT JOB

39. What do you think makes you to stay in your present job ?

Lack of opportunity and resources to go abroad ( )
You like your job ( )
Your family circumstances ( )
Your loyalty/affinity to your boss ( )
Your loyalty/affinity to your Organisation ( )
Good future prospects in your present assignment ( )

PERMANENT ADDRESS

40. Please give your permanent address for any future communication/contacts :

.....

.....

.....

EDITING PRIMARY DATA

The completed questionnaires and schedules must be carefully checked and edited for errors. This is quite a difficult job and requires a great deal of skill and experience. While editing primary data the following considerations need attention :

1. The data should be complete.
2. The data should be consistent.
3. The data should be accurate.
4. The data should be homogeneous.

1. **Editing for completeness.** The editor should see that each schedule and questionnaire is complete in all respects, *i.e.*, answer to each and every question has been furnished. If some questions have not been answered and those questions are of vital importance, the informants should be contacted again either personally or through correspondence. It may happen that in spite of best efforts a few questions remain unanswered. In such questions, the editor should mark 'Not reported' or simply N.R. in the space provided for answers and if the questions are of vital importance then the schedule or questionnaire could be dropped.

2. **Editing for consistency.** While editing the data for consistency, the editor should see that answers to questions are not contradictory in nature. If there are mutually contractory answers, he should try to obtain the correct answers either by referring back the questionnaire or by contacting, wherever possible, the informant in person. For example, if amongst others, two questions in a questionnaire are : (*a*) Are you married ? (*b*) State the number of children you have, and the reply to the former question is "no" and to the latter 'Three', then there is contradiction and it should be clarified.

3. **Editing for accuracy.** The reliability of the inferences drawn depend basically on the correctness of information. If the information supplied is wrong, inferences can never be valid. It is, therefore, necessary for the editor to see that the information is accurate in all respects. However, this is one of the most difficult tasks of the editor. If the inaccuracy is due to arithmetical errors, it can be easily detected and corrected. But if the cause of inaccuracy is faulty information supplied it may be difficult to verify it, for example, information relating to income, age, sales, etc.

4. **Editing for homogeneity.** By homogeneity we mean that all the questions have been understood in the same sense by different respondents. The editor must check various questions carefully. If some informants have given monthly income, others annual income and still others weekly income or even daily income, no comparison can be made. Similarly, if some persons have given the basic income whereas others the total income, no comparison is possible. The editor should check that information supplied by the various people is homogeneous and uniform.

It should be noted that these days computer is extensively being used to edit data. Various computer techniques have been developed to identify "outliers"—responses which are greatly different from the majoritv of the responses. Many outliers result from clerical error, recording, transcription or from false information provided by the respondent.

PROBLEMS

1-A. Answer the following questions, each questions carries **one** mark :

(*i*) Define primary data.
(*ii*) What is secondary data ?
(*iii*) Give few sources of secondary data.
(*iv*) What is meant by questionnaire ?
(*v*) Why pre-testing of questionnaire is desirable ?
(*vi*) What is editing ?
(*vii*) How many questions should a questionnaire contain ?
(*viii*) Is primary source more reliable than secondary source ?
(*ix*) Name few sources of business data.
(*x*) Statistics are dangerous in the hands of the inexpert.

1-B. Answer the following questions, each question carries **four** marks :

(*i*) Distinguish between primary and secondary data.
(*ii*) What are the methods of collecting primary data ?
(*iii*) Briefly explain the characteristics of a good questionnaire.
(*iv*) What precautions should be taken while using secondary data.
(*v*) Distinguish between structured and unsourced questionnaire.
(*vi*) Describe the various steps that are taken in conducting a statistical investigations.
(*vii*) Discuss the merits and limitations of collecting primary data through questionnaire.

2. Distinguish clearly between internal and external data. Give examples to illustrate the distinction between the two.

3. Distinguish between primary and secondary data. Discuss the various methods of collecting primary data. Indicate the situation in which each of these methods should be used.
[*MBA, HPU, 2002; MBA (HCA), DU., 2002; MBA, UP Tech. Univ., 2003*]

4. Distinguish between primary and secondary data. What precautions would you take before using data from a secondary source ?

5. "It is never safe to take published statistics at their face value without knowing their meaning and limitation." Elucidate this statement by enumerating and briefly explaining the various points which you consider before using any published statistics. Illustrate your answer with example wherever possible.

6. Discuss the validity of the statement : "A secondary source is not as reliable as primary source."

7. Explain what precautions must be taken while drafting a questionnaire in order that it may be really useful. Illustrate your answer giving suitable examples.

8. As the personnel manager of a particular firm you want to determine the effect of pecuniary and non-pecuniary incentives on workers' efficiency. Draft a suitable questionnaire.

9. Describe the different methods of collecting data indicating the merits and demerits of each of them. Which method is suitable to the following types of studies ?
   (i) Enquiry by a Research Organisation into the living conditions of the workers of cotton textile mills of Bombay.
   (ii) Study of the buying habits of the people in regard to washing powder like Surf, Lux, Nirma, etc.
   (iii) Enquiry to the food situation by a committee appointed by the Government of India.

10. Are the following statements true, false, or a combination of truth and false, or a combination of truth and falsehood ?
   (a) Bias is not undesirable if it contributes to the reporting of results that the investigator has anticipated.
   (b) Inaccurate responses to questionnaire create no serious problem for the investigator as long as they result from inability to reply correctly rather than from bias.
   (c) Interviews introduce more bias than does the use of questionnaire.
   (d) 'True or false' or 'Yes or No' questions should not be used in questionnaires unless only one of the two answers is possible.
   (e) Open questions are more difficult than most other types to tabulate.

11. (a) A manufacturing organisation has selling branches in each large town in the country. It makes 6 kinds of articles which are sold both in retail and wholesale by the branches. The Head Office wishes to plan a sales campaign based on the past sales and likely future demand. Design a questionnaire for the collection of the necessary data and draft instructions for completing the questionnaire.
   (b) What are the sources of secondary data ? Explain some uses and limitations of secondary sources of data.
   [*MBA, Osmania, 2002*]
   (c) Compare and contrast the questionnaire and interview techniques of collecting data. Which technique is more reliable and why ? [*MBA, TU, Kathmandu, 2002*]

12. Distinguish between the following :
   (a) a schedule and a questionnaire,
   (b) primary and secondary data, and
   (c) survey and an experiment.

13. In constructing a questionnaire or a schedule, the primary steps are design, pre-test and editing. Describe briefly each of these.

14. In the following situations indicate whether a sample or a census should be taken and explain why :
   (i) A car manufacturer wants to obtain data on customer performances with respect to size of cars.
   (ii) A firm employing 1,200 persons wants to determine the acceptability of subscribing to a new employee insurance programme.
   (iii) The AGCR office wants to obtain data on the proportion of income-tax returns that contain arithmetic mistakes.
   (iv) A researcher wants to find out efficiency of an executive development programme.

15. In the following situations, which method of data collection—self, enumeration-personal interview or telephone interview—would you select and why ? You should also keep in mind the cost of the method, response rate and the time necessary to obtain the information, as well as other relevant factors.
   (i) Consumer acceptance of new TV model before it is placed on the market.
   (ii) Data on percentage expenditure incurred on education by class IV employees of State.
   (iii) Information of the adequacy of the social security measure and the changes to be made therein.
   (iv) The determination of national ranking of an international management convention.
   (v) Data on malaria cases during 2003-04.

16. In the following set of questions find at least one fault in each. Also suggest an improved rewording of the questions :
   (i) How many tubes of toothpaste did you purchase in the last six months ?
   (ii) Do you agree that too much money is being spent on entertainment by various ministries ?
   (iii) Does the name National Panasonic come to your mind while buying a record player ?
   (iv) Is it a waste of money spend heavily on defence (strongly agree, agree, undecided, disagree, strongly disagree) ?
   (v) What inspired you to join this company ?

17. Define 'secondary data'. State their chief sources and point out the dangers involved in their use and precautions necessary before using them.

18. Distinguish between primary and secondary data. Give a brief account of the chief methods of collecting primary data and bring out their merits and defects.
19. (*a*) Describe the requirements of a good questionnaire.
    (*b*) Construct a suitable questionnaire containing not more than twenty questions pertaining to the effectiveness of an EDP programme attended by your subordinates.
20. It is required to collect information on the economic conditions of textile workers in Bombay. Suggest a suitable method for collection of primary data. Draft a suitable questionnaire of about ten questions for collecting this information. Also suggest how will you proceed to carry out statistical analysis of the information collected.
21. "Data collected in census are automatically free from errors." Discuss the validity of this statement.
22. (*a*) Examine critically any two method of collecting primary data.
    (*b*) Distinguish clearly between structured and unstructured questionnaires. Construct a suitable questionnaire containing not more than ten questions pertaining to 'Consumer Survey' on Maruti Cars.
23. What are the basic sources of business data ? What precautions would you take while using secondary data ?
24. (*a*) What are the methods of collecting primary data ? Can field and laboratory experiments act as sources of business data ? Explain briefly.
    (*b*) What precautions must be taken while drafting a suitable questionnaire ? Give examples. [*MBA, Rohilkhand Univ., 2007*]
25. What are the essentials of a good questionnaire ? Draft a suitable questionnaire containing not more than 20 questions to find out the efficacy of MBA programme. [*MBA, Osmania Univ., 2006*]
26. Suppose you are in charge of conducting a socio-economic survey of the Taxi drivers in a city. Prepare a suitable questionnaire in this connection.
27. You are required to collect data on the extent and nature of graduate unemployment in urban areas in Nepal, using the sample survey method. How do you proceed ? [*MBA, TU, 2008*]

SCREEN ACTORS GUILD

Read the following case carefully and answer the questions given in the end :

Evaluating a marketing research project

The Screen Actors Guild shared with all unions the goal of ensuring its 29,000 members a living wage and job security. But lately the Guild had become increasingly interested, as well, in the accuracy and honesty of its members dramatic roles. A driving force behind this trend was Louise Garrity, who was a vice president in the Guild and who also headed its Women's Conference Committee.

Last year Ms. Garrity appeared before the Federal Communications Committee (FCC) and at a meeting of televisions network producers. Her purpose in these appearances was to discuss such issues as the media's image of women and minorities, TV reruns, the prime-time access rule, and the "family viewing time" rule.

Recently the Women's Conference Committee of the Guild, under Ms. Garrity's leadership, launched a national survey of television viewers' opinions. Although the survey was particularly concerned with how women were portrayed on television, it was also designed to touch on other areas.

"This is not just a question of women's image," Ms. Garrity was quoted. "We want to be able to go before the FCC and network producers and studio people and writers—especially the writers—and say that x amount of people in this or that area like this or don't like that. We can't go on feelings. We need facts, and input from around the country, to perform a better service as entertainers. It's a matter of projecting *truth*."

An attitude questionnaire was designed (see Exhibit 1). In 192 communities the editors of newspapers, Sunday magazine supplements, and TV sections of newspapers were invited to complete the questionnaire. Readers were invited to complete the questionnaire, adding any personal comments they wished to make, and send it directly to the Screen Actors Guild. (The Guild's address was to be provided when the questionnaire was published.)

QUESTIONS

1. Evaluate the research design and methodology used in this project. 30 Marks
2. Evaluate the questionnaire in Exhibit 1.

EXHIBIT—1 SCREEN ACTORS GUILD ATTITUDES QUESTIONNAIRE

Occupation _____ Male _____ Female _____ City and State : _____

Age level :

Under 12 _____
Under 18 _____
Under 25 _____
Under 35 _____
Under 45 _____
Under 55 _____
55 and over _____

Education :

Grade school level _____
High school level _____
Some college _____
College degree _____

1. Do you think television influences the way you live your own life ?
   (a) Mode of dress _____ Yes _____ No _____ Undecided.
   (b) Mode of conduct _____ Yes _____ No _____ Undecided.
   (c) Products you buy _____ Yes _____ No _____ Undecided.
   (d) Attitudes about minorities _____ Yes _____ No _____ Undecided.
   (e) Attitudes about women _____ Yes _____ No _____ Undecided.
2. Do you think that the images of women presented on TV are truthful and believable ? _____ Yes _____ No _____ Undecided.
3. Do you like the women you see on TV ?
   _____ Yes _____ No _____ Undecided.
4. Do you feel that women are abused and ridiculed by media more often than not ? _____ Yes _____ No _____ Undecided.
5. Do you feel the relationships and roles on TV shows mirror women's life-styles ? _____ Yes _____ No _____ Undecided.
6. Do you feel that the media encourage young girls to aspire to useful and meaningful roles in society ? _____ Yes _____ No _____ Undecided.
7. Do you identify with the women in daytime soap operas ?
   _____ Yes _____ No _____ Undecided
8. Do you think there is hostility between women as portrayed in television commercials ?
   _____ Yes _____ No _____ Undecided.
9. Do you think commercials portray women's total identity and happiness as depending on the use of the product ?
   _____ Yes _____ No _____ Undecided.
10. Do you think sex is overused to products ?
    _____ Yes _____ No _____ Undecided.
11. Do you feel that women's news items are given equal time and serious consideration ?
    _____ Yes _____ No _____ Undecided.
12. Would you like to see more women in leading roles on television programs other than comedies and variety and talk shows ?
    _____ Yes _____ No _____ Undecided.
13. Are you aware of the small number of women appearing in dramatic shows ? _____ Yes _____ No _____ Undecided.
14. Would you like to see women appearing on TV in positions of authority ?
    (*a*) Presenting national news ?
    _____ Yes _____ No _____ Undecided.
    (*b*) Moderators of game shows ?
    _____ Yes _____ No _____ Undecided.
    (*c*) Hosts of talk shows and children's programmes ?
    _____ Yes _____ No _____ Undecided.

(*d*) Spokeswomen for national products ?
_____ Yes _____ No _____ Undecided.

(*e*) Voiceovers (the voice you hear off camera) ?
_____ Yes _____ No _____ Undecided.

(*f*) Narrators of documentaries ?
_____ Yes _____ No _____ Undecided.

(*g*) Actresses portraying women in professions ?
_____ Yes _____ No _____ Undecided.

15. Do you see any change in minority representation ?

(*a*) Black _____ Yes _____ No _____ Undecided.
(*b*) Mexican _____ Yes _____ No _____ Undecided.
(*c*) Asian _____ Yes _____ No _____ Undecided.
(*d*) Indian _____ Yes _____ No _____ Undecided.
(*e*) Other _____ Yes _____ No _____ Undecided.

16. Do you feel the image of minorities is accurately represented ?
_____ Yes _____ No _____ Undecided.

17. What is your preference for prime-time evening programming ?
Marks in order of preferences :
Drama _____ Documentaries _____ Variety _____ Comedy _____
Game shows _____ Sports _____ .

18. Do you feel television generally is fantasy or fact ?
Fantasy _____ Fact _____ .

19. Regarding reruns of TV series in prime time, do you think there are
_____ Too few _____ Too many _____ Just enough _____ Undecided.

20. Do you think the public should have some say in how many shows are rerun ?
_____ Yes _____ No _____ Undecided. *[MBA, Sukhadia Univ., 2007]*

*****

3

Presentation of Data

INTRODUCTION

After the data have been collected, the next step is to present them in some suitable form. The need for proper presentation arises because of the fact that statistical data in their raw form almost defy comprehension. When data are presented in easy to-read form, it can help the reader to acquire knowledge in much shorter period of time and also facilitate statistical analysis, Presentation can take two basic forms : (*i*) Statistical Table, and (*ii*) Statistical Chart.

A statistical table is presentation of numbers in a logical arrangement, with some brief explanation to show what they are. However, before, tabulating data it is often necessary to first classify them. A statistical chart or a graph is a pictorial device for presenting data. The present chapter has been divided into three main parts to enable greater clarity : (*A*) Classification of data, (*B*) Tabulation of data, (C) Charting data.

(A) CLASSIFICATION OF DATA

After collection and editing of data an important step towards processing the data is classification. Classification is the grouping of related facts into different classes. Facts in one class differ from those of another class with respect to some characteristics called a basis of classification. Sorting facts on one basis of classification and then on another basis is called cross-classification. This process can be repeated as many times as there are possible basis of classification. Classification of data is a function very similar to that of sorting letters in a post office. It is well known that the letters collected in a post office are sorted into different lots on a geographical basis, *i.e.*, in accordance with their destinations as Mumbai, Kolkata, Kanpur, Jaipur, etc. They are then put in separate bags, each containing letters with a common characteristic, *viz.*, having the same destination. Classification of statistical data is comparable to the sorting operation. The process of classification gives prominence to important information gathered while dropping unnecessary details facilitates comparison and enables a statistical treatment of the material collected.

Types of Classification

Broadly, the data can be classified on the following four basis :

(*i*) Geographical, *i.e.*, area-wise, *e.g.*, cities, districts, etc.

(*ii*) Chronological, *i.e.*, on the basis of time.

(*iii*) Qualitative, *i.e.*, according to some attributes.

(*iv*) Quantitative, *i.e.*, in terms of magnitudes.

(*i*) **Geographical classification.** In geographical classification data are classified on the basis of geographical or locational differences between the various items. For example, when we present the production of sugar cane, wheat, rice, etc., for various States, this would be called geographical classification.

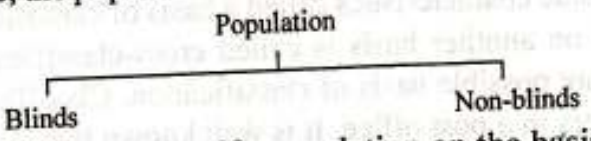
Geographical classifications are usually listed in alphabetical order for easy reference. Items may also be listed by size to emphasise the important areas as in ranking the States by population. Normally in reference tables the first approach is followed and in summary tables the second approach is followed.

(*ii*)**Chronological classification.** When data are observed over a period of time, the type of classification is known as chronological classification. For example, the sales figures of a company are given below :

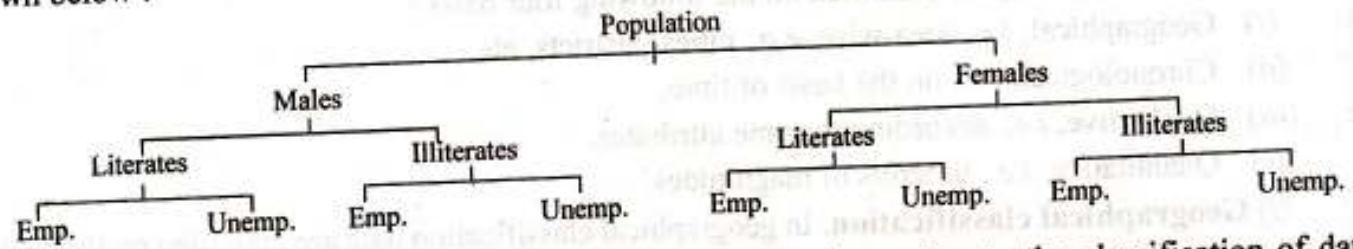
Year	*Sales (Rs. lakh)*	*Year*	*Sales (Rs. lakh)*
2000-01	18810	2005-06	46725
2001-02	23601	2006-07	45724
2002-03	23816	2007-08	50117
2003-04	32435	2008-09	53900
2004-05	39343	2009-10	61795

Time series are usually listed in chronological order normally starting with the earliest time period. When the major emphasis falls on the most recent events, a reverse time order may be used.

(*iii*) **Qualitative classification.** In qualitative classification, data are classified on the basis of some attribute or quality such as sex, colour of hair, literacy, religion, etc. The point to note in this type of classification is that the attribute under study cannot be measured: one can only find out whether it is present or absent in the units of the population under study. For example, if the attribute under study is blindness, we may find out how many persons are blind in a given population. It is not possible to measure the degree of blindness in each case. Thus when only one attribute is studied, two classes are formed; one possessing the attribute and the other not possessing it. This type of classification is known as simple classification. For example, the population under study may be divided into two categories as follows :



In a similar manner, we may classify population on the basis of sex, *i.e.*, males and females, or literacy, *i.e.*, literates and illiterates, and so on. This type of classification where only two classes are formed is also called twofold or dichotomous classification. If instead of forming only two classes we further divide the data on the basis of some attribute or attributes so as to form several classes, the classification is known as manifold classification. For example, we may first divide the population into males and females on the basis of the attribute, 'sex'; each of these classes may be further subdivided into 'literates and illiterates' on the basis of the attribute 'literacy'. Further classification can be made based on some other attribute, say, employment. The type of manifold classification described here is shown below :



(*iv*) **Quantitative classification.** Quantitative classification refers to the classification of data according to some characteristics that can be measured, such as height, weight, income, sales, etc. For example, the workers of a factory may be classified according to wages as follows :

Monthly wages (Rs.)	*No. of workers*	*Monthly wages (Rs.)*	*No. of workers*
4000-4500	50	5500-6000	360
4500-5000	200	6000-6500	90
5000-5500	260	6500-7000	40
		Total	1,000

In this type of classification, there are two elements, namely (*i*) the *variable*, *i.e.*, the monthly wage in the above example, and (*ii*) *frequency*, *i.e.*, the number of workers in each class. There were 50 workers having income between Rs. 4000 and Rs. 4500, 200 workers having income between Rs. 4500 and Rs. 5000, and so on. The quantitative classification gives birth to a frequency distribution.*

Variable. A frequency distribution refers to data classified on the basis of some variable that can be measured such as prices, wages, age, number of units produced or consumed. The term, 'variable' refers to the characteristic that varies in amount or magnitude in a frequency distribution. A variable may be either continuous or discrete (also called discontinuous). A continuous variable is capable of manifesting every conceivable fractional value within the range of possibilities, such as the height or weight of persons or the weight of a product. Thus, as a student grows, say, from 90 cm to 150 cm, his height passes through all values between these lines. On the other hand, a discrete variable is that which can try only be finite "jumps" and cannot manifest every conceivable fractional value. For instance, the number of rooms in a house can only take certain values as 1, 2, 3, etc. Similarly, the number of employees and number of machines in an establishment are discrete variables. Generally speaking, continuous data are obtained through measurements, while discontinuous data are derived by counting. Series which can be described by a continuous variable are called continuous series. Series represented by a discrete variable are called discrete series. The following are two examples of discrete and continuous frequency distributions :

No. of Children	*No. of Families*
0	10
1	400
2	800
3	700
4	250
5	150
6	50
	Total 2,360

(*a*) Discrete Frequency Distribution

Age (years)	*No. of Employees*
20—25	10
25—30	15
30—35	40
35—40	45
40—45	26
45—50	4
	Total 140

(*b*) Continuous Frequency Distribution

Although the theoretical distinction between continuous and discrete variables is clear and precise, in practical statistical work it is only an approximation. The reason is that even the most precise instruments of measurement can be used only to finite number of places. Thus from a practical viewpoint continuous series can never be expected to flow continuously with one measurement touching another without any break in actual observations.

Formation of a Frequency Distribution

The process of preparing this type of distribution is very simple. We have just to count the number of times a particular value is repeated which is called the frequency of that class. In order to facilitate counting, prepare a column of 'tally'. In another column, place all possible values of the variable from

* The word 'distribution' refers to the way in which the observations are distributed in different classes.

the lowest to the highest. Then, put a bar (vertical line) opposite the particular value to which it relates. To facilitate counting, blocks of five bars are prepared and some space is left in between each block. We finally count the number of blocks and bars corresponding to each value of the variable and place it in the column of frequency. The process shall be clear from the following example of the number of refrigerators sold on 22 working days by a leading agency house :

23, 30, 20, 26, 30 30, 20, 23, 40., 40, 26, 20, 23, 40, 28, 26, 23, 30,
40, 28, 28, 30.

FREQUENCY DISTRIBUTION OF THE NUMBER OF REFRIGERATORS SOLD

No. of Refrigerators	*Tally Bars*	*Frequency No. of days*
20	𝍸	3
23	𝍸 |	4
26	𝍸	3
28	𝍸	3
30	𝍸	5
40	𝍸 |	4
	Total	22

The table clearly shows that on 3 days 20 refrigerators were sold each day, on 4 days 23 refrigerators were sold each day, etc.

This method of classifying helps in condensing the data only where values are largely repeated, otherwise there will be hardly any condensation. In order to make the series more compact so that its characteristics can be easily studied, data may be classified according to class-intervals.

Classification according to Class-Intervals

This type of classification is most popular in practice. The following technical terms are important when data are classified according to class-intervals :

(*i*) **Class limits.** The class limits are the lowest and the highest values that can be included in the class. For example, take the class 20—40. The lowest value of this class is 20 and the highest 40. The two boundaries of a class are known as the lower limit and upper limit of the class. The lower limit of a class is the value below which there can be no value in that class. The upper limit of a class is the value above which no value can belong to that class. Of the class 70—89, 70 is the lower limit and 80 is the upper limit, *i.e.*, in this class, there can be no value which is less than 70 or more than 89. Similarly, if we take the class 90—109, there can be no value in that class which is less than 90 or more than 109.

(*ii*) **Class-intervals.** The span of a class, that is, the difference between the upper limit and the lower limit, is known as class-interval. For example, in the class 20—40, the class interval is 20 (*i.e.*, 40 minus 20). The size of the class-interval is determined by the number of classes and the total range in the data.

(*iii*) **Class frequency.** The number of observations corresponding to the particular class is known as the frequency of that class or the class frequency. In the illustration given on page 41, the frequency of the class 5000-6000 is 50 which implies that there are 50 employees having income between Rs. 5000 and Rs. 6000. If we add together the frequencies of all individual classes, we obtain the total frequency. Thus, in the same problem, the total frequency of the six classes is 550 which means that in all there are 550 employees whose income has been studied.

(*iv*) **Class mid-point.** It is the value lying half-way between the lower and upper class limits of a class-interval. Mid-point of a class is ascertained as follows :

$$\text{Mid-point of a class} = \frac{\text{Upper limit of the class} + \text{Lower limit of the class}}{2}$$

For the purpose of further calculations in statistical work the mid-point of each class is taken to represent that class.

There are two methods of classifying the data according to class-intervals, namely (*a*) 'exclusive' method, and (*b*) 'inclusive' method.

(*a*) **'Exclusive' method.** When the class-intervals are so fixed that the upper limit of one class is the lower limit of the next class, it is known as the 'exclusive' method of classification. The following data are classified on this basis:

Income (Rs.)	*No. of Employees*	*Income (Rs.)*	*No. of Employees*
5000-6000	50	8000-9000	150
6000-7000	100	9000-10000	40
7000-8000	200	10000 and above	10
			Total 550

It is clear that the 'exclusive' method ensures continuity of data inasmuch as the upper limit of one class is the lower limit of the next class. Thus, in the above example, there are 50 employees whose income is between Rs. 5000 and Rs. 5999.99. An employee who is getting exactly Rs. 6000 would be included in the class 6000-7000. This method is widely followed in practice. However, it is confusing to a layman who has no knowledge of statistics. For example, if a questionnaire includes an observation asking the respondent the number of times he visits the Super Bazar in a month and he is required to tick one of the categories: 5-10 and 10-15, a person who visits the Super Bazar 10 times may find it difficult to decide whether to put the tick in the space against the class 5-10 and 10-15. In the absence of any specific instructions, some people may tick the class 5-10 while others 10-15. Hence, whenever this method is used it is necessary to give clear instructions in the questionnaire. However, the reader should note that if class-intervals are given like 0-10, 10-20, etc., it is always presumed that upper limit is exclusive, *i.e.*, an observation exactly equal to the upper limit is not included in that class.

(*b*) **'Inclusive' method.** Under the 'inclusive' method of classification, the upper limit of one class is included in that class itself. The example given on the next page illustrates this method:

Income (Rs.)	*No. of Employees*	*Income (Rs.)*	*No. of Employees*
5000-5999	50	8000-8999	150
6000-6999	100	9000-9999	40
7000-7999	200	10000-10999	10
			Total 550

In the class 5000-5999, we include employees whose income is between Rs. 5000 and Rs. 5999. If the income of an employee is exactly Rs. 6000 he is included in the next class. The above example makes it clear that there is no confusion here of the type we find under the 'exclusive' method. We may have classes like 5000-5999.5 or 5000-5999.9, and so on.

It should be noted that both the inclusive and exclusive methods give us the same class frequencies, although the class-intervals are apparently different in the two cases. In the above example, in case of exclusive method the class interval is 100 whereas in case of inclusive method the class-interval is 99. However, 99 is not the correct class-interval. The correct class-interval is 100. It is because whenever 'inclusive' method is used for equal class-intervals, the class-interval is obtained by taking the difference between the two upper limits.

Principles of Classification

It is difficult to lay down any hard and fast rules for classifying the data as the type of classification. However, the following general considerations may be borne in mind for ensuring meaningful classification of data:

(1) The number of classes should preferably be between 5 and 15. However, there is no rigidity about it. The classes can be more than 15 depending upon the total number of observations in the data and the details required, but they should not be less than five because in that case the classification may not reveal the essential characteristics.

Struges suggested the following formula for determining the approximate number of classes:

$k = 1 + 3.322 \log N.$

k = The approximate number of classes.

N = Total number of observations.

log = The ordinary logarithm to the base of 10.

However, the precise number of classes to be used for a given variable depends upon personal judgment and other considerations such as the details required, the ease of calculation of further statistical work, etc.

(2) As far as possible one should avoid odd values of class-intervals, *e.g.*, 3, 7, 11, 26, 39, etc. Preferably, one should have class-intervals of either five or multiples of five like 10, 20, 25, 100, etc. The reason is that the human mind is accustomed more to think in terms of certain multiples of 5, 10 and the like. However, where the data necessitate a class-interval of less than 5 it can be any value between 1 and 4.

(3) The starting point, *i.e.*, the lower limit of the first class, should either be zero or 5 or multiple of 5. For example, if the lowest value of the data is 63 and we have taken a class-interval of 10, then the first class should be 60-70, instead of 63-73. Similarly, if the lowest value of the series is 76 and the class-interval is 5 then the first class should be 75 to 80 rather than 76 to 81.

(4) To ensure continuity and to get correct class-interval we should adopt 'exclusive' method of classification. However, where 'inclusive' method has been adopted it is necessary to make an adjustment to determine the correct class-interval and to have continuity. The adjustment consists of finding the difference between the lower limit of the second class and the upper limit of the first class, dividing the difference by two, subtracting the value so obtained from all lower limits and adding the value to all upper limits. This can be expressed in the formula as follows :

$$\text{Correction factor} = \frac{\text{Lower limit of the 2nd class} - \text{Upper limit of the 1st class}}{2}$$

How the adjustment is made when data are given by inclusive method can be seen from the following example :

Monthly Wages (Rs.)	*No. of Workers*	*Monthly Wages (Rs.)*	*No. of Workers*
5000-5999	5	8000-8999	18
6000-6999	10	9000-9999	12
7000-7999	15	10000-10999	4

To adjust the class limits, we take the difference between 6000 and 5999 which is one. By dividing it by two we get 1/2 or 0.5. This (0.5) is called the correction factor. Deduct 0.5 from the lower limits of all classes and add 0.5 to upper limits. The adjusted classes would then be as follows :

Monthly Wages (Rs.)	*No. of Workers*	*Monthly Wages (Rs.)*	*No. of Workers*
4999.5-5999.5	5	7999.5-8999.5	18
5999.5-6999.5	10	8999.5-9999.5	12
6999.5-7999.5	15	9999.5-10999.5	4

(5) Whenever possible all classes should be of the same size. If intervals are not of uniform width, it is difficult to make meaningful comparison between classes. At times, however, extreme observations may require the inclusion of so many class-intervals that the frequency distribution will become unwieldy. The observations are then classified as follows: below 200, 200-400, 400-600, 600-800, 800 and above. These classes are called *open-end classes* and distribution is known as an *open-end frequency distribution*. When frequency distribution is being employed as a technique of presentation only, open-end classes do not seriously reduce its usefulness as long as only a few observations fall in these classes. However, use of the frequency distribution for purposes of further mathematical computation is not helpful because a mid-point value, which can be used to represent the class, cannot be determined for an open-end class.

It may be noted that the frequency table, like other types of data presentation, is always constructed to serve some specific purpose. The technical requirements outlined above must be supplemented by sound subjective judgement if proper and useful frequency distributions are to be formed.

Illustration 1. The profits (in lakhs of rupees) of 30 companies for the year 1999-2000 are given below :

20, 22, 35, 42, 37, 42, 48, 53, 49, 65, 39, 48, 67, 18, 16, 23, 37, 35, 49, 63, 65, 55, 45, 58, 57, 69, 25, 29, 58, 65.

Classify the above data taking a suitable class-interval.

Solution: Let us determine the suitable class-interval with the help of the following formula:

$$i = \frac{\text{Range}}{1 + 3.322 \log N}$$

Range = (69–16) = 53, $N = 30$

$$i = \frac{53}{(1 + 3.322 \times 1.4771)} = \frac{53}{1 + 4.91} = \frac{53}{5.91} = 8.97 \text{ or } 9.$$

Since values like 3, 7, 9, etc., should be avoided and therefore, we will take 10 as the class-interval and the first class as 15–25.

FREQUENCY DISTRIBUTION OF THE PROFITS

Profits (Rs. lakhs)	*Tally Bars*	*No. of Companies*
15–25	𝍸	5
25–35	𝍸	2
35–45	𝍸 ‖	7
45–55	𝍸 |	6
55–65	𝍸	5
65–75	𝍸	5
		Total 30

Illustration 2. Present the following data of the marks of 60 applicants who were given a certain test for the purpose of selection to a post :

41	17	83	63	55	92	60	58	70	06
67	82	33	44	57	49	34	73	54	63
36	52	32	75	60	33	09	79	28	30
42	93	43	80	03	52	57	67	84	64
63	11	35	28	10	23	08	41	60	32
72	53	92	88	62	55	60	33	40	57

Take first class as 0–9.

Solution : **FREQUENCY DISTRIBUTION OF THE MARKS OF 60 APPLICANTS**

Marks	*Tally Bars*	*Frequency*
0–9	𝍸	4
10–19	𝍸	3
20–29	𝍸	3
30–39	𝍸 𝍸	10
40–49	𝍸 𝍸	7
50–59	𝍸 𝍸	9
60–69	𝍸 𝍸	11
70–79	𝍸	5
80–89	𝍸	5
90–99	𝍸	3
		Total 60

Illustration 3. The data given below relate to the sales and advertisement expenditure of 20 companies. You are required to form a bivariate frequency distribution with class interval 62 to 64, 64 to 66, and so on and 115 to 125, 125 to 135 and so on.

Company	*Sales (Rs. Lakhs)*	*Adv. Exp. (Rs. Lakhs)*	*Company*	*Sales (Rs. Lakhs)*	*Adv. Exp. (Rs. Lakhs)*
1	170	70	11	163	70
2	135	65	12	139	67
3	136	65	13	122	63
4	137	64	14	134	68
5	148	69	15	140	67
6	124	62	16	132	69
7	117	65	17	120	66
8	128	70	18	148	68
9	143	71	19	120	67
10	129	62	20	152	67

Solution : As per the requirements of the question, the data are to be divided into five classes according to the advertisement expenditure and six classes according to the sales.

For tabulating the information in appropriate cells, first, the row to which the advertisement expenditure (say, X) should belong is determined. Afterwards on a consideration of the sales (say, Y) the column in which it should be included is determined. The tabulation is recorded by tally bars. Thus the two-way table shall be prepared as follows :*

TWO-WAY FREQUENCY TABLE SHOWING SALES AND ADVERTISEMENT EXPENDITURE OF 20 COMPANIES

Adv. Exp. (X) \ *Sales (Y)*	115–125	125–135	135–145	145–155	155–165	165–175	Total
62–64	II (2)	I (1)	–	–	–	–	3
64–66	I (1)	–	III (3)	–	–	–	4
66–68	II (2)	–	II (2)	I (1)	–	–	5
68–70	–	II (2)	–	II (2)	–	–	4
70–72	–	I (1)	I (1)	–	I (1)	I (1)	4
Total	5	4	6	3	1	1	20

*The figure in brackets denote the frequency corresponding to each cell, advertisement expenditure incurred and six classes according to the sales. There will be thus $5 \times 6 = 30$ cells.

(B) TABULATION OF DATA

One of the simplest and most revealing devices for summarizing data and presenting them in meaningful fashion is the statistical table.* A table is a systematic arrangement of statistical data in columns and rows. Rows are horizontal arrangement, whereas columns are vertical ones. The purpose of a table is to simplify the presentation and to facilitate comparisons. The simplification results from the clear-cut and systematic arrangement, which enables the reader to quickly locate desired information. Comparison is facilitated by bringing related items of information close together.

Parts of a Table

The various parts of a table may vary from case to case depending upon the given data. But a good table must contain at least the following parts :

1. Table number
2. Title of the table
3. Caption
4. Stub
5. Body of the table
6. Headnote
7. Footnote

1. **Table number.** Each table should be numbered. There are different practices with regard to the place where this number is to be given. The number may be given either in the centre at the top above the title or in the side of the table at the top or at the bottom of the table on the left-hand side. However, if space permits the table number should be given in the centre. Where there are many columns, it is also desirable to number each column so that easy reference to it is possible.

2. **Title of the table.** Every table must have a suitable title. The title is a description of the contents of the table. A complete title has to answer the questions *what, where* and *when* in that sequence. In other words,

(*a*) What precisely are the data in the table (*i.e.*, what categories of statistical data are shown)?

(*b*) Where the data occurred (*i.e.*, the precise geographical, political or physical area covered)?

(*c*) When the data occurred (*i.e.*, the specific time or period covered by the statistical material on the table)?

The title should be clear, brief and self-explanatory. However, clarity should not be sacrificed for the sake of brevity. Long title cannot be read as promptly as short title, but at times they may have to be used for the sake of clarity. The title should be so worded that it permits one and only one interpretation. It should be in the form of a series of phrases rather than complete sentences. *Its lettering should be the most prominent of any lettering in the table.*

3. **Caption.** Caption refers to the column headings. It explains what the column represents. It may consist of one or more column headings. Under a column-heading there may be sub-heads. The caption should be clearly defined and placed at the middle of the column. If the different columns are expressed in different units, the unit should be specified along with the captions. As compared with the main part of the table the caption should be shown in smaller letters. This helps in saving space.

4. **Stub.** As distinguished from caption, stubs are the designation of the rows or row headings. They are at the extreme left and perform the same function for the horizontal rows or numbers in the table as the column headings do for the vertical columns or numbers. The stubs are usually wider than column headings but should be kept as narrow as possible without sacrificing precision and clarity of statements.

---

* A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrase and statements to the form of titles, headings and notes to make clear the full meaning of data and their origin.

(B) TABULATION OF DATA

One of the simplest and most revealing devices for summarizing data and presenting them in meaningful fashion is the statistical table.* A table is a systematic arrangement of statistical data in columns and rows. Rows are horizontal arrangement, whereas columns are vertical ones. The purpose of a table is to simplify the presentation and to facilitate comparisons. The simplification results from the clear-cut and systematic arrangement, which enables the reader to quickly locate desired information. Comparison is facilitated by bringing related items of information close together.

Parts of a Table

The various parts of a table may vary from case to case depending upon the given data. But a good table must contain at least the following parts :

1. Table number
2. Title of the table
3. Caption
4. Stub
5. Body of the table
6. Headnote
7. Footnote

1. **Table number.** Each table should be numbered. There are different practices with regard to the place where this number is to be given. The number may be given either in the centre at the top above the title or in the side of the table at the top or at the bottom of the table on the left-hand side. However, if space permits the table number should be given in the centre. Where there are many columns, it is also desirable to number each column so that easy reference to it is possible.

2. **Title of the table.** Every table must have a suitable title. The title is a description of the contents of the table. A complete title has to answer the questions *what, where* and *when* in that sequence. In other words,

(*a*) What precisely are the data in the table (*i.e.*, what categories of statistical data are shown)?

(*b*) Where the data occurred (*i.e.*, the precise geographical, political or physical area covered)?

(*c*) When the data occurred (*i.e.*, the specific time or period covered by the statistical material on the table)?

The title should be clear, brief and self-explanatory. However, clarity should not be sacrificed for the sake of brevity. Long title cannot be read as promptly as short title, but at times they may have to be used for the sake of clarity. The title should be so worded that it permits one and only one interpretation. It should be in the form of a series of phrases rather than complete sentences. *Its lettering should be the most prominent of any lettering in the table.*

3. **Caption.** Caption refers to the column headings. It explains what the column represents. It may consist of one or more column headings. Under a column-heading there may be sub-heads. The caption should be clearly defined and placed at the middle of the column. If the different columns are expressed in different units, the unit should be specified along with the captions. As compared with the main part of the table the caption should be shown in smaller letters. This helps in saving space.

4. **Stub.** As distinguished from caption, stubs are the designation of the rows or row headings. They are at the extreme left and perform the same function for the horizontal rows or numbers in the table as the column headings do for the vertical columns or numbers. The stubs are usually wider than column headings but should be kept as narrow as possible without sacrificing precision and clarity of statements.

---

* A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrase and statements to the form of titles, headings and notes to make clear the full meaning of data and their origin.

5. **Body of the table.** The body of the table contains the numerical information. This is the most vital part of the table. Data presented in the body arranged according to descriptions and classifications of the captions and stubs.

6. **Headnote.** It is a brief explanatory statement applying to all or a major part of the material in the table, and is placed below the title entered and enclosed in brackets. It is used to explain certain points relating to the whole table that have not been included in the title nor in the captions or stubs. For example, the unit of measurement is frequently written as the headnote, such as "in thousands" or "in million tonnes", or "in crores", etc.

7. **Footnote.** Anything in a table which the reader may find difficult to understand from the title, captions and stubs should be explained in footnotes. If footnotes are needed, they are placed directly below the body of the table. Footnotes are used for four main purposes :

(*a*) To point out any exceptions as to the basis of arriving at the data, for example, sales recorded at 'delivered price' for others. Any heterogeneity in the data recorded must be disclosed to avoid wrong conclusions.

(*b*) Any special circumstances affecting the data, for example, strike, lock-out, fire, etc.

(*c*) To clarify anything in the table.

(*d*) To give the source in case of secondary data. The reference to the source should be complete in itself, for example, if the data are obtained from some periodical, its name, date of publication, page number, table number, etc., should be mentioned so that if the user wishes to check the data from the original source, he will know where to look for the information.

There are various systems of identifying the footnotes. One is numbering them consecutively with small number ^{1}, ^{2}, ^{3}, or letters *a, b, c, d.* Another system identifies the first footnote with one star (*), second footnote with two stars (**), third footnote with three stars (***), and so on. Sometimes instead of star another sign (dagger) is used. However, where several footnotes are required, it is more convenient to use small numbers like ^{1}, ^{2}, ^{3}, ... etc.

Review of the Table

Before a table is released it should be reviewed for form, content, validity and clerical accuracy. It is difficult for the person preparing the table to make a thoroughly satisfactory checks on all the four aspects. The person, who has prepared the table, might have done his best, but he can hardly review it objectively. He should, if possible, get his work reviewed by some experienced person.

In case of a summary table, the reviewer should ask himself the following questions to determine whether or not table is satisfactory :

(1) Does the title clearly state what is in the table ?

(2) Are all the entries pertinent ?

(3) Is there unity of subject-matter ?

(4) Are the classifications arranged so as to focus attention on the main comparisons ?

(5) Are the data arranged so as to emphasise important points ?

(6) Does the table include adequate interpretative figures such as totals, percentages and averages ?

(7) Are there notations about peculiarities of the data ?

(8) Is the source properly stated ?

(9) Is the table in proper form, so that it presents an attractive appearance ?

Types of Tables

Tables may broadly be classified into two categories :

1. Simple and complex tables ; and
2. General purpose and special purpose (or summary) tables.

1. **Simple and Complex Tables.** The distinction between simple and complex tables is based upon the number of characteristics studied.

In a simple table only one characteristic is shown. Hence, this type of table is also known as one-way table. In a complex table, on the other hand, two or more characteristics are shown. Such tables are more popular in practice because they enable full information to be incorporated and facilitate a proper consideration of all related facts. When two characteristics are shown, a table is known as two-way table or double tabulation. When three characteristics are shown in a table, this type of tabulation is known as treble tabulation. When four or more characteristics are simultaneously shown, it is a case of manifold tabulation. The following examples will illustrate the distinction between simple and complex tables :

(*i*) *Simple table or one-way table.* In this type of table only one characteristic is shown. This is the simplest of tables. The following is the illustration of such a table :

NUMBER OF EMPLOYEES IN AN ORGANISATION ACCORDING TO AGE GROUP

Age (in years)	*No. of Employees*
Below 25	50
25—35	67
35—45	43
45—55	15
55 and above	5
	Total 180

(*ii*) *Two-way table.* Such a table shows two characteristics and is formed when either the stub or the caption is divided into two coordinate parts. The example given on page 47 illustrates the nature of such a table :

NUMBER OF EMPLOYEES IN AN ORGANISATION ACCORDING TO AGE AND SEX

Age (in years)	*Employees*		*Total*
	Males	*Females*	
Below 35	32	18	50
25—35	40	27	67
35—45	25	18	43
45—55	10	5	15
55 & above	5	—	5
Total	112	68	180

(*iii*) *Higher order table.* When three or more characteristics are represented in the same table, such a table is called higher order table. The need for such a table arises when we are interested in presenting a number of characteristics simultaneously. While constructing such a table it is necessary to first establish an order of precedence among the attributes or characteristics sought to be classified having regard to their relative importance.

It should be remembered that as the number of characteristics represented increases, the table becomes more and more confusing and as such normally not more than four characteristics should be represented in the same table. Where more than four characteristics are to be represented we can have more than one table depicting relationship between different attributes.

2. **General Purpose and Special Purpose Tables.** General purpose tables, also known as the reference tables or repository tables, provide information for general use or reference. They usually contain detailed information and are not constructed for specific discussion. In other words, these tables serve as a repository of information and are arranged for easy reference. Tables published by governmental agencies are mostly of this kind, such as the tables contained in the *Statistical Abstract of the Indian Union,* detailed tables contained in the census reports, etc. Such tables tell facts which are not for particular discussion. When such table are used by a researcher, they are usually placed in the appendix of the report for easy reference.

Special purpose tables, also known as summary or analytical tables, provide information for prticular discussion. They show relationship between different groups of figures. When attached to a report they are found in the body of the text. These tables are also called derivative tables since they are often derived from general tables. Thus the large detailed tables in the census records of the Government of India are general purpose tables. When such data are used, they are ordinarily taken from the general purpose tables and presented as special purpose tables, which emphasise the relation the user wishes to stress. A special purpose table should be designed in such a way that a reader may easily refer to the table for comparison, analysis or emphasis concerning the particular discussion.

Illustration 4. In a sample study about the coffee habits in two towns, following data were observed:

Town *X*
52% persons were males,
25% were coffee drinkers, and
16% were male coffee drinkers

Town *Y*
55% persons were males,
28% were coffee drinkers, and
18% were male coffee drinkers.

Tabulate the above observations. *(MBA, HPU, 2006)*

Solution.

TABLE SHOWING PERCENTAGE OF COFFEE DRINKERS

Attributes	*Town X*			*Town Y*		
	Males	*Females*	*Total*	*Males*	*Females*	*Total*
Coffee drinkers	16	9	25	18	10	28
Non-coffee drinkers	36	39	75	37	35	72
Total	52	48	100	55	45	100

Illustration 5. Present in a tabular form with suitable captions, etc., for the information contained in the following:

"In 2000, out of a total of 2,000 workers of a factory 1,500 workers were members of a trade union. The number of women employed was 150 of which 128 did not belong to a trade union. In 2005, the number of trade union workers increased to 1,620 of which 1,582 were men. On the other hand, the number of non-union workers fell down to 448 of which 318 were men. In 2010, there were on the payrolls of the factory 2,200 workers of whom 2,000 belonged to a trade union. Of all the employees in 2010, 200 were women of whom only 25 did belong to a trade union."

Solution.

TABLE SHOWING TRADE UNION MEMBERSHIP

	2000			2005			2010		
Category	*Trade Union members*	*Non-members of T.U.*	*Total*	*Trade Union members*	*Non-members of T.U.*	*Total*	*Trade Union members*	*Non-members of T.U.*	*Total*
Men	1,478	372	1,850	1,582	318	1,900	1,825	175	2,000
Women	22	128	150	38	130	168	175	25	200
Total	1,500	500	2,000	1,620	448	2,068	2,000	200	2,200

(C) CHARTING DATA

One of the most convincing and appealing ways in which data may be presented is through charts. Evidence of this can be found in the financial pages of newspapers, journals, advertisements, etc. Pictorial presentation helps in quick understanding of the data. As the number and magnitude of figures increases, they become more confusing and their analysis tends to be more strenuous. A picture is said to be worth 10,000 words, *i.e.*, through pictorial presentation data can be presented in an interesting form. Not only this, charts have greater memorizing effect as the impressions created by them last much longer than those created by the figures.

A chart can take the shape of either a diagram or a graph. For the sake of clarity we will discuss them under two separate heads :

(*i*) Diagram, and (*ii*) Graphs.

(*i*) Diagrams

For representing data diagrams are more commonly used than graphs. However, before discussing different types of diagrams it would be worthwhile to consider some general rules for constructing diagrams.

General Rules for Constructing Diagrams

The following general rules should be observed while constructing diagrams :

1. *Title.* Every diagram must be given a suitable title. The title should convey in as few a words as possible the main idea that the diagram is intended to portray. However, the brevity should not be secured at the cost of clarity or omission of essential details. The title may be given either at the top of the diagram or below it.

2. *Proportion between width and height.* A proper proportion between the height and width of the diagram should be maintained. If either the height or width is too short or too long in proportion, the diagram would give an ugly look. While there are no fixed rules about the dimensions, convenient standard as suggested by Lutz in the book entitled *Graphic Presentation,* may be adopted for general use. It is known as "Root-two", that is ratio of 1 (short side) to 1.414 (long side). Modifications wherever necessary may be made to accommodate a diagram in the space available.

3. *Selection of appropriate scale.* The scale showing the values should be in even numbers or in multiples of five or ten, *e.g.*, 25, 50, 75 or 20, 40, 60. Odd values like 1, 3, 5, 7 should be

avoided. No rigid rules can be laid down about the selection of appropriate scale. The given data and the required size of diagram are the guiding factors. The scale should specify the size of the unit and what it represents, for example, "millions of tonnes", "number of persons in thousands", "units produced in lakhs", etc. *All lettering should be easily readable without turning the chart sidewise.*

4. *Footnotes.* In order to clarify certain points about the diagram, footnotes may be given at the bottom of the diagram.

5. *Index.* An index illustrating different types of lines or different shades, colours, should be given so that the reader can easily make out the meaning of the diagram.

6. *Neatness and cleanliness.* Diagrams should be absolutely neat and clean.

7. *Simplicity.* Diagrams should be as simple as possible so that the reader can understand their meaning clearly. For the sake of simplicity, it is important that too much material should not be loaded in a single diagram otherwise it may become too confusing and prove useless. Several simple charts are much better and more effective than one or two complex ones which present the same material in a confusing way.

Types of Diagrams

In practice, a very large variety of diagrams are in use and new ones are constantly being added. It would be outside the scope of this text to deal exhaustively with the subject and as such only more frequently used diagrams are discussed. For the sake of convenience and simplicity different types of diagrams are divided under the following heads :

I. One-dimensional diagrams, *e.g.*, bar diagrams.
II. Two-dimensional diagrams, *e.g.*, rectangles, squares and circles.
III. Pictograms and cartograms.

Each of these types is discussed below in detail :

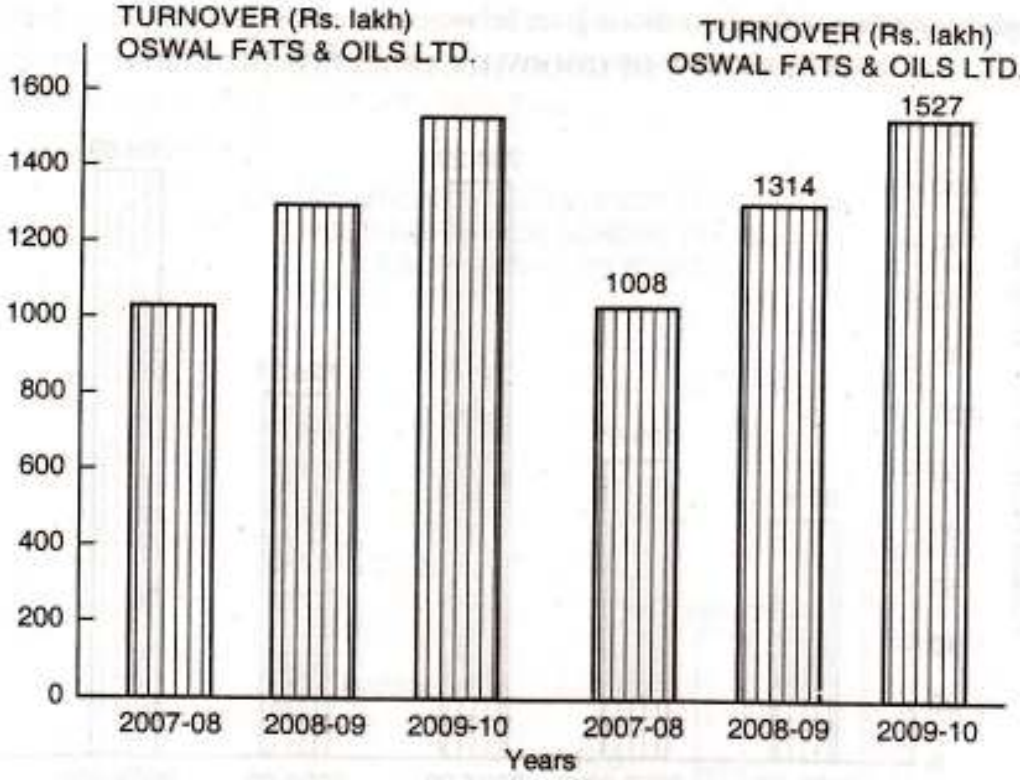
I. One-dimensional or Bar Diagrams

Bar diagrams are the most common type of diagrams used in practice. A bar is a thick line whose width is shown merely for attention. They are called one-dimensional because it is only the length of the bar that matters and not the width. When the number of observations is large, lines may be drawn instead of bars to economise space. Special merits of bar diagram are the following :

(*i*) They are readily understood even by those unaccustomed to reading charts or those who are not chart-minded.
(*ii*) They possess the outstanding advantage that they are the simplest and the easiest to make.
(*iii*) When a large number of observations are to be compared, they are the only form that can be used effectively.

Points to be kept in mind while constructing Bar Diagrams

(*i*) The width of the bars should be uniform throughout the diagram.
(*ii*) The gap between one bar and another should be uniform throughout.
(*iii*) Bars may be either horizontal or vertical. The vertical bars should be preferred because they give a better look and also facilitate comparison.
(*iv*) While constructing the bar diagrams, it is desirable to write the respective figure at the end of each bar so that the reader can know the precise value without looking at the scale. This is particularly important when the scale is too narrow; for example, 1 cm on paper may represent 10 crore people. The following two diagrams would clarify the difference :



It is clear from the above two diagrams that from the left one it is difficult to read precise values whereas the right side diagram makes it clear.

Types of Bar Diagrams

Bar diagrams are of the following types:

(*a*) Simple bar diagrams
(*b*) Subdivided bar diagrams*
(*c*) Multiple bar diagrams
(*d*) Precentage bar diagrams
(*e*) Deviation bars
(*f*) Broken bars

(a) Simple Bar Diagrams

A simple bar diagram is used to represent only one variable. For example, the figures of sales, production, population, etc., for various years may be shown by means of a simple bar diagram. Since the bars are of the same width and only the length varies, it becomes very easy for the reader to study the relationship. Simple bar diagrams are very popular in practice. However, an important limitation of such diagrams is that they can present only one classification or one category of data. For example, while presenting the population for the last five decades, one can only depict the total population in the simple bar diagrams and not its sex-wise distribution.

Illustration 6. The funds flow of Goodwill India Ltd. from 2005-06 to 2009-10 are given below :

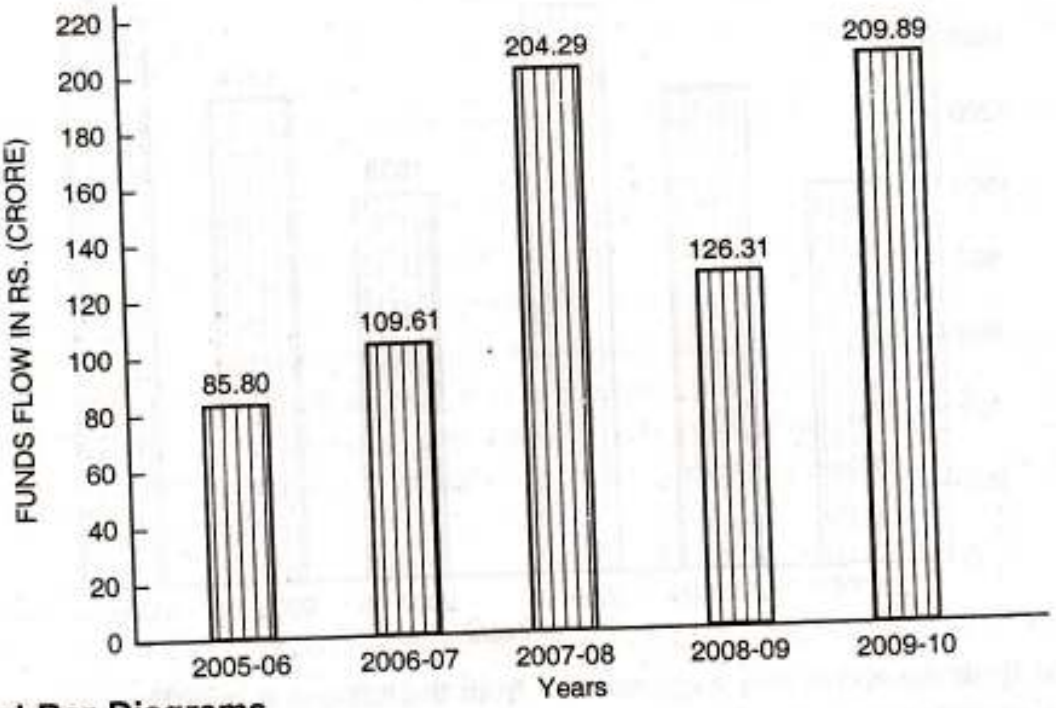
Year	*Funds Flow (Rs. crores)*
2005-06	85.80
2006-07	109.61
2007-08	204.29
2008-09	126.31
2009-10	209.89

Represent this data by a suitable bar diagram.

*Such diagrams are also known as component bar diagrams.

Solution. The simple bar diagram of the above data is given below.

FUNDS FLOW OF GOODWILL INDIA LTD. (Rs. Crore)



(*b*) Subdivided Bar Diagrams

These diagrams are used to represent various parts of the total. For example, the number of employees in various departments of a company may be represented by a subdivided bar diagram. While constructing such a diagram, the various components in each bar should be kept in the same order. A common and helpful arrangement is that of presenting each bar in the order of magnitude from the largest component at the base of the bar to the smallest at the end. To distinguish between the different components, it is useful to use different shades or colours. Index or key should be given explaining these differences. Sub-divided bar diagrams can be vertical as well as horizontal.

Subdivided bar diagrams should not be used where the number of components is more than 10 or 12, for, in that case, the diagram would be overloaded with information which cannot be easily compared and understood.

The component bar diagrams can be used to represent either the absolute data or distribution ratios such as percentage distribution ratios is, in fact, an excellent method for presenting a set of distribution ratios diagrammatically.*

Illustration 7. Represent the following data by subdivided bar diagram.

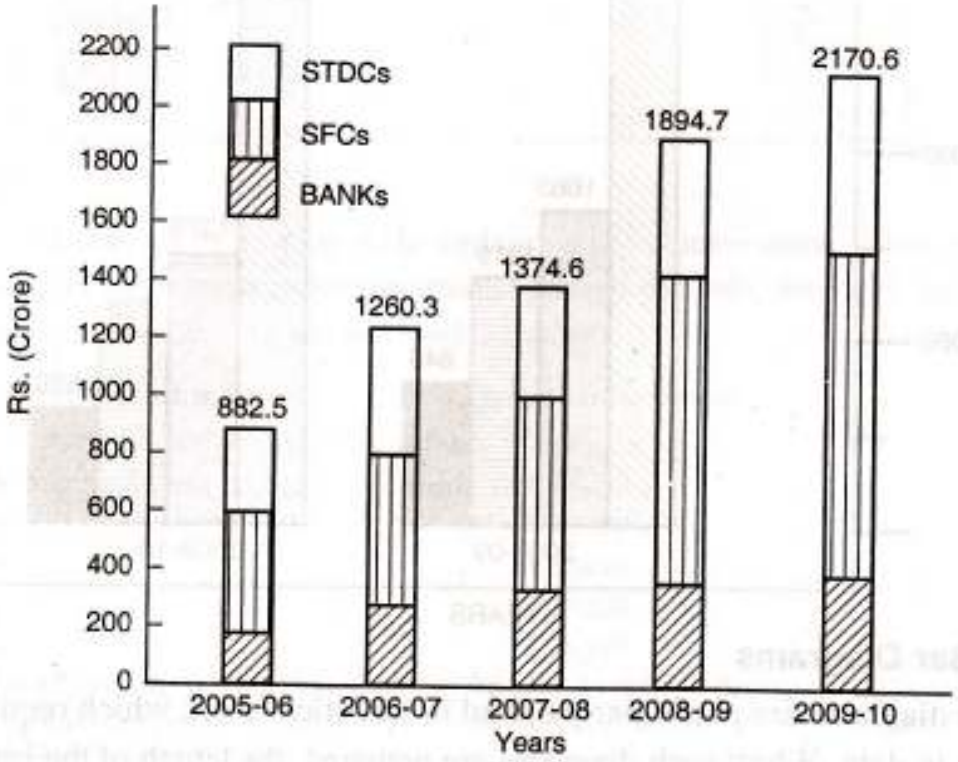
INSTITUTIONWISE ASSISTANCE SANCTIONED UNDER REFINANCE SCHEME (Rs. Crore)

Year	*Banks*	*SFCs*	*STDCs*	*Total*
2005-06	233.8	365.3	283.4	882.5
2006-07	301.8	484.7	473.8	1260.3
2007-08	303.2	668.6	402.8	1374.6
2008-09	365.3	992.8	536.6	1894.7
2009-10	416.4	1067.4	686.8	2170.6

*The other alternatives for this purpose are the relative pie diagram and the relative component line chart. The latter can be used only in cases where the classification is chronological. When the number of time period is not large, the bar chart is undoubtedly superior to these diagrammatic methods.

Solution. Since we have to show three different variables, subdivided bar diagram will be more appropriate. In order to prepare such a diagram, bar is to be drawn of the total of all the three heads for each year and then it is to be subdivided in three heads. Thus for 2005-06, total is 882.5; for 2006-07, 1260.3, etc.

DISTRIBUTION OF ASSISTANCE SANCTIONED
UNDER REFINANCE SCHEME (Rs. Crore)
FROM 2005-06 TO 2009-10



(c) Multiple Bar Diagrams

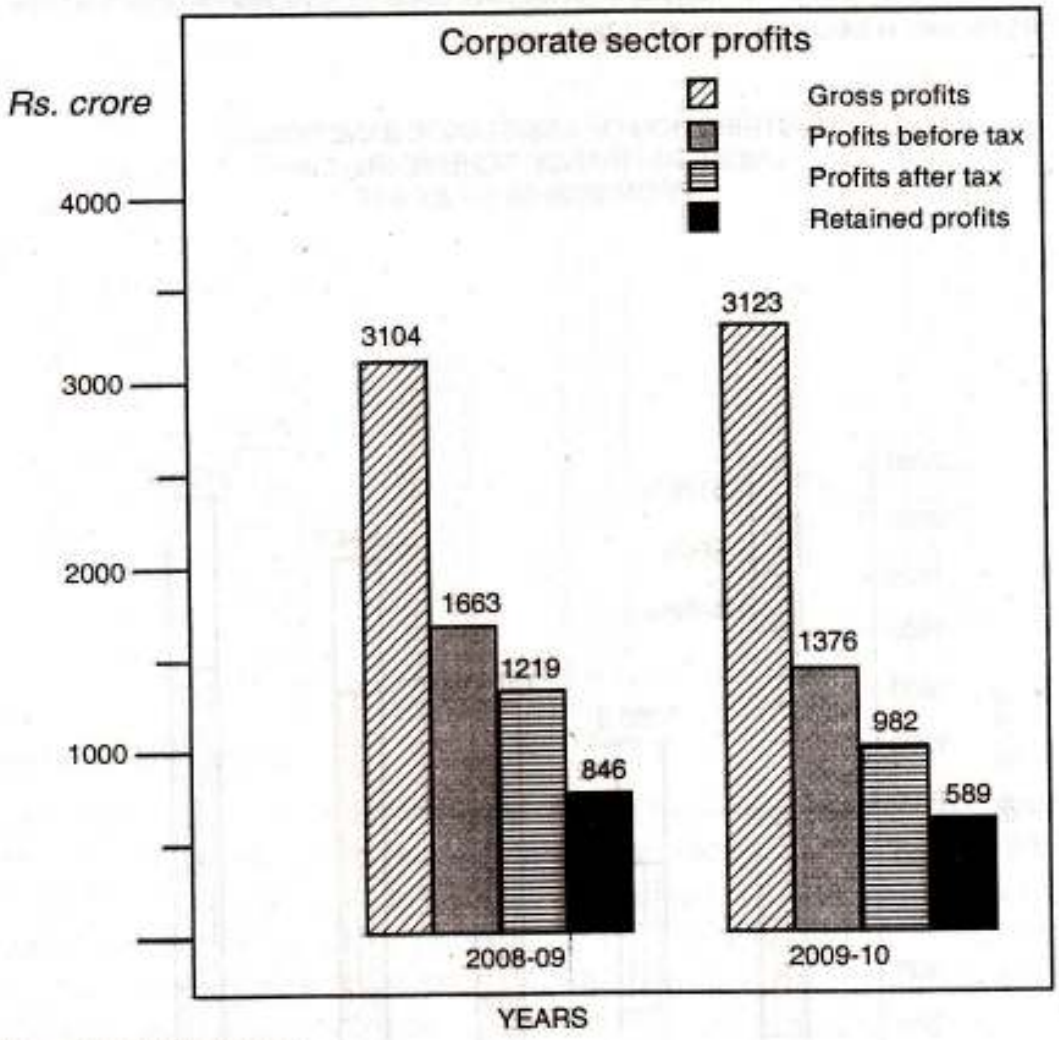
In multiple bar diagram two or more sets of interrelated data are represented. The technique of drawing such a diagram is the same as that of simple bar diagram. The only difference is that since more than one phenomenon is represented, different shades, colours, dots, or crossings are used to distinguish between the bars.

Illustration 8. Represent the following data by multiple bar diagram :

CORPORATE SECTOR PROFITS (Rs. crore)

	2008-09	2009-10
Gross profits	3104	3123
Profits before tax	1663	1376
Profits after tax	1219	982
Retained profits	846	589

Solution. The multiple bar diagram of the above data is given below :



(*d*) Percentage Bar Diagrams

Percentage bar diagrams are particularly useful in statistical work which requires the portrayal of relative changes in data. When such diagrams are prepared, the length of the bars is kept equal to 100 and segments are cut in these bars to represent the components (percentages) of an aggregate.

(*e*) Deviation Bars

Deviation bars are popularly used for representing net quantities—excess or deficit, *i.e.*, net profit, net loss, net exports or imports, etc. Such bars can have both positive and negative values. Positive values are shown above the base line and negative values below it. The following illustration would explain this type of diagram :

Illustration 9. The following are the figures of sales and net profits of public sector units over the last three years. Represent the data by a suitable diagram.

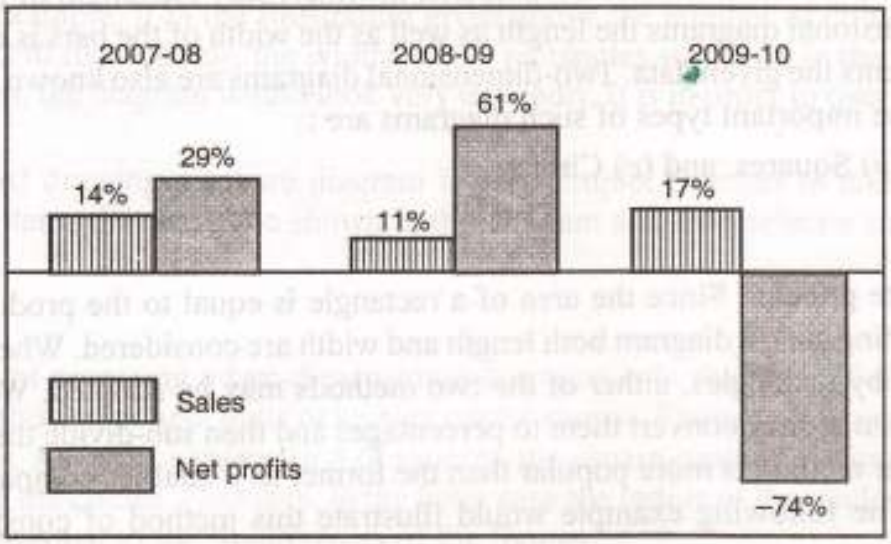
(% change over previous year)

Year	*Sales*	*Net profits*
2007-08	14%	29%
2008-09	11%	61%
2009-10	17%	–74%

Solution. The above data can best be represented by deviation bars.

Sales & Net Profits of public sector units

(% change over previous year)



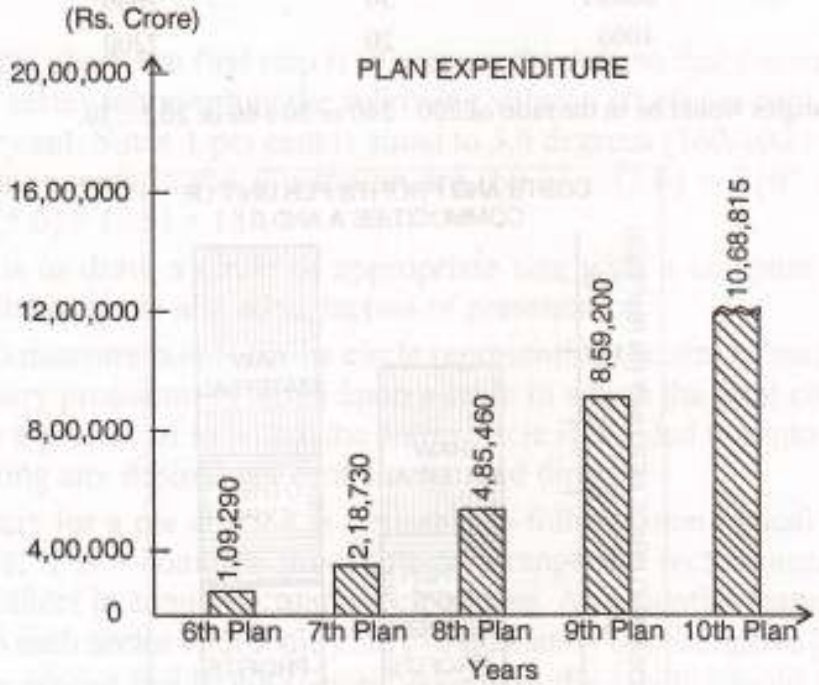
(f) Broken Bars

In certain type of data there may be wide variations in values—some values may be very small, other very large. In order to gain space for the smaller bars of the data, the large bars may be broken.

Illustration 10. Represent the following data by a suitable diagram :

Five-Year Plan	*Actual public sector outlay (Rs. crore)*
6th	1,09,290
7th	2,18,730
8th	4,85,460
9th	8,59,200
10th	10,68,815

Solution. 10th Plan expenditure is more than double of 8th plan. In order to gain space we have broken the bar for 10th Plan. Otherwise, the length of this bar would have almost 2 times that of the bar for 9th Plan and the diagram would have occupied a lot of space and given an ugly look.



* Estimate.

II. Two-dimensional Diagrams

As distinguished from one-dimensional diagrams in which only the length of the bars is taken into account, in two-dimensional diagrams the length as well as the width of the bars is considered. Thus the area of the bar represents the given data. Two-dimensional diagrams are also known as *surface diagrams* or *area diagrams*. The important types of such diagrams are :

(*a*) Rectangles, (*b*) Squares, and (*c*) Circles.

(*a*) Rectangles

This form is quite popular. Since the area of a rectangle is equal to the product of its length and width, while constructing such a diagram both length and width are considered. When two sets of figures are to be represented by rectangles, either of the two methods may be adopted. We may represent the figures as they are given or may convert them to percentages and then sub-divide the length into various components. The latter method is more popular than the former as it enables comparison to be made on a percentage basis. The following example would illustrate this method of constructing rectangular diagrams :

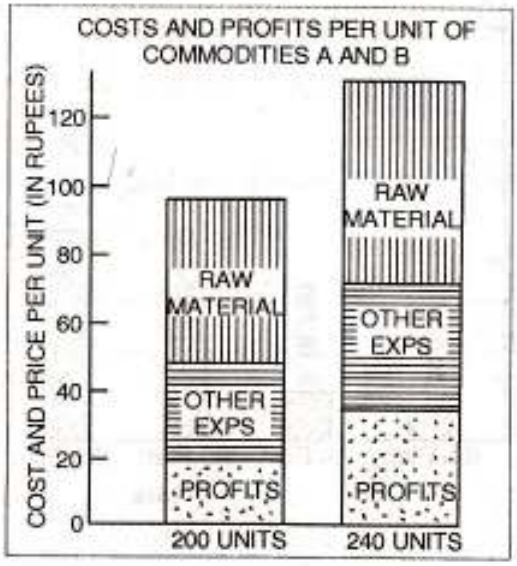
Illustration 11. Present the following data by a rectangular diagram :

	Commodities	
	A	*B*
Price per unit of commodity (Rs.)	100	120
Quantity sold	200	240
Cost of raw materials used (Rs.)	10000	12000
Other costs (Rs.)	6000	9600
Profit (Rs.)	4000	7200

Solution. Let us calculate the cost of raw materials, other expenses and profits per unit.

	Commodity A 200 units		*Commodity B 240 units*	
	Total (Rs.)	*Per unit (Rs.)*	*Total (Rs.)*	*Per unit (Rs.)*
Cost of raw material	10000	50	12000	50
Other expenses	6000	30	9600	40
Profit	4000	20	7200	30

The widths of the rectangles would be in the ratio of 200 : 240 or 50 : 60 or 20.5 : 30.



(b) Squares

The rectangular method of diagrammatic presentation is difficult to use where the values of items vary widely. For example, if in the illustration given above the number of units sold of commodities *A* and *B* are 20 and 240 respectively, the widths of the rectangles would be in the ratio of 5 : 60 or 1 : 12. If this ratio is taken, the diagram would look very unwiedly. It is in order to overcome this difficulty that squares are used.

The method of drawing a square diagram is very simple. One has to take the square-root of the values of various items that are to be shown in the diagram and then select a suitable scale to draw the squares.

(c) Circles

Another way of preparing a two-dimensional diagrams is in the form of circles. In such diagrams both the total and the component parts or sectors can be shown. The area of a circle is proportional to the square of its radius. As in the construction of squares, the square-roots of various figures are worked out while constructing the circles. However, in the latter case the radius of the circles (rather than the side of squares) are proportional to the square-roots of the figures.

Circles can be used in all those cases in which squares are used. However, in both these types of diagrams it is difficult to judge the relative magnitude with precision.

Circles are difficult to compare and as such they are not very popular in statistical work. When it is necessary to use circles, they should be compared on an area basis rather than on a diameter basis as the diameter basis is very misleading. Compared to rectangles, circles are more difficult to construct and interpret.

Pie Diagram

This type of diagram enables us to show the partitioning of a total into component parts. A very common use of the pie chart is to represent the division of a sum of money into its components. For example, the entire circle, or pie, may represent the budget of a family for a month and the sections may represent portions of the budget alloted to rent, food, clothing, and so on. Similarly, through a pie diagram we can show how a rupee spent by a firm is distributed over various heads such as wages, raw materials, administration expenses, etc.

The pie chart is so called because the entire graph looks like a pie, and the components resemble slices cut from it.

In constructing a pie chart, the first step is to prepare the data so that the various component values can be transposed in a series representing the following values : (*i*) 60 per cent, (*ii*) 25 per cent, (*iii*) 10 per cent, and (*iv*) 5 per cent. Since 1 per cent is equal to 3.6 degrees (360/100 = 3.6), the corresponding values of the four components in the illustration are (60.0) × (3.6) = 216° ; (25.0) × (3.6) = 90° ; (10.0) × (3.6) = 36° ; (5.0) × (3.6) = 18°.

The second step is to draw a circle of appropriate size with a compass. The size of the radius depends upon the available space and other factors of presentation.

The third step is to measure points on the circle representing the size of each sector with the help of a protractor. The ordinary protractor is based upon a scale in which the total circle is 360 degree, but it is possible to purchase a protractor in which the entire circle is divided not into 360 but 100 equal parts so that angle representing any desired per cent can be read directly.

In laying out sectors for a pie chart, it is desirable to follow some logical arrangement, pattern or sequence. For example, it is a common procedure to arrange the sectors according to size, with the largest at the top and others in sequence running clockwise. An essential feature of the pie chart is the careful identification to each sector with some kind of explanatory or descriptive label. If there is sufficient room the labels can be placed inside the sectors; otherwise the labels should be placed in continuous positions outside the circle, usually with an arrow pointing to the appropriate sector.

Illustration 12. The following funds in assistance were sanctioned and disbursed during 2002-03 to 2009-10 by a leadi financial institution.

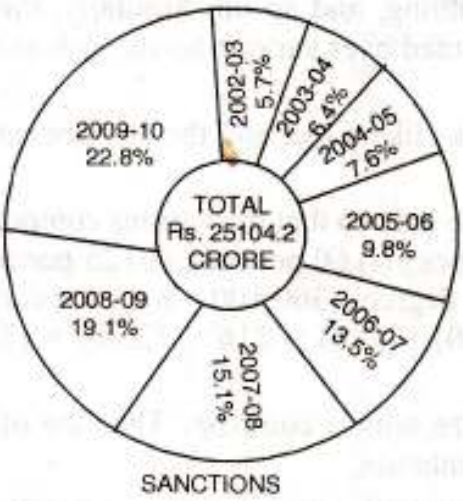
Year	*Amount sanctioned (Rs. crore)*	*Amount disbursed (Rs. crore)*
2002-03	1444.3	1066.1
2003-04	1607.0	1339.5
2004-05	1905.5	1582.4
2005-06	2449.1	1961.5
2006-07	3394.6	2293.4
2007-08	3784.4	2787.0
2008-09	4791.5	3258.0
2009-10	5727.8	4041.6

Represent the data by a pie diagram.

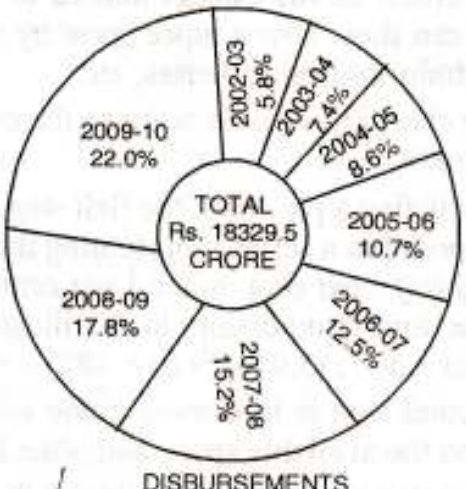
Solution. Convert the given data into percentages and then prepare two pie diagrams and divide them into segments.

Year	*Amount sanctioned (Rs. crore)*	*%*	*Amount disbursed (Rs. crore)*	*%*
2002-03	1444.3	5.7	1066.1	5.8
2003-04	1607.0	6.4	1339.5	7.4
2004-05	1905.5	7.6	1582.4	8.6
2005-06	2449.1	9.8	1961.5	10.7
2006-07	3394.6	13.5	2293.4	12.5
2007-08	3784.4	15.1	2787.0	15.2
2008-09	4791.5	19.1	3258.0	17.8
2009-10	5727.8	22.8	4041.6	22.0
Total	25104.2	100.0	18329.5	100.0

Composition of Trends in Assistance Sanctioned and Disbursed during the year 2002-03 to 2009-10.



SANCTIONS



DISBURSEMENTS

Limitations of Pie Diagrams

Pie diagrams are less effective than bar diagrams for accurate reading and interpretation, particularly when series are divided into a large number of components or the difference among the components is very small. It is generally inadvisable to attempt to portray a series of more than five or six categories by means of a pie chart. If, for example, there are eight, ten or more categories it may be very confusing to differentiate the relative values portrayed especially when the several small sectors are of approximately the same size. This type of diagram, although frequently used, appears upon comparison inferior to simple bar diagram the divided bar diagram or a group of curves.

III. Pictograms and Cartograms

(i) Pictograms

Pictograms, also known as picturegrams, are very popularly used in presenting statistical data. They are not abstract presentations such as lines or bars, but really depict the kind of data we are dealing with. Pictures are attractive and easy to comprehend and as such this method is particularly useful in presenting statistics to the layman. When pictograms are used, data are represented through a pictorial symbol that is carefully selected. The picture symbol should be self-explanatory in nature, *i.e.*, it should represent clearly the phenomena. For example, if the increase in number of buses on road is shown over a period of time the appropriate symbol would be a bus.

Illustration 13. The following table gives the production of tea in India by a leading company :

Years	2006	2007	2008	2009	2010
Production (Million kgs)	421	561	587	645	660

Represent the data by a pictogram.

Solution. For representing the above data by a pictogram we will use the symbol of a cup.

PICTOGRAM

PRODUCTION OF TEA	
YEAR	(million kgs)
2006	421
2007	561
2008	587
2009	645
2010	660

Merits. (1) As compared with other types of diagrams, pictograms have a greater attraction value and, therefore, where the attention of masses is to be drawn such as in exhibitions, fairs, etc., they are very popularly used. They stimulate interest in the information being presented.

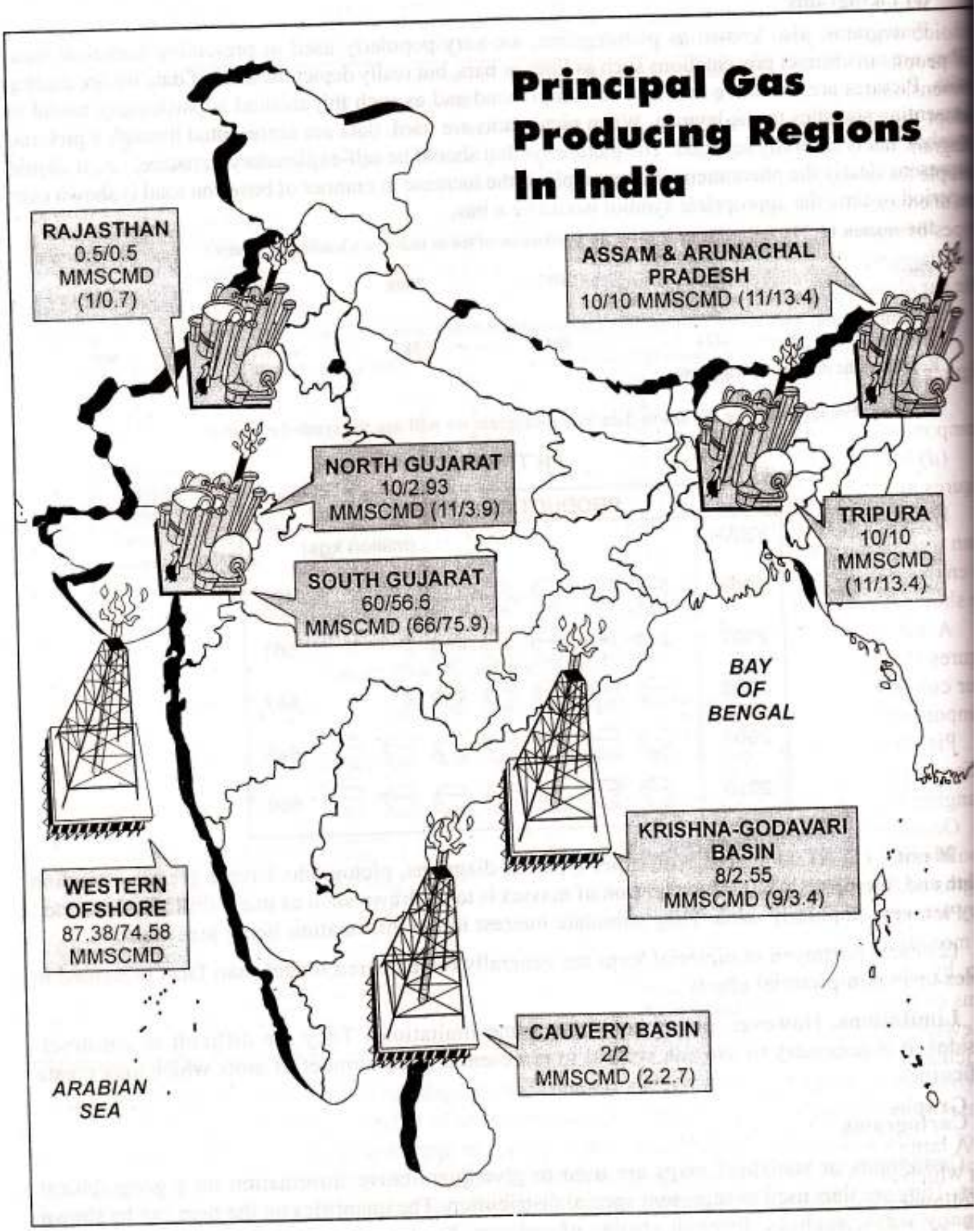
(2) Facts portrayed in pictorial form are generally remembered longer than facts presented in tables or in non-pictorial charts.

Limitations. However, pictograms have some limitations. They are difficult to construct. Besides, it is necessary to use one symbol to represent a fixed number of units which may create difficulties.

(ii) Cartograms

Cartograms or statistical maps are used to give quantitative information on a geographical basis. They are thus used to represent special distribution. The quantities on the map can be shown in many ways, such as, through shades of colours, by dots, by placing pictograms in each geographical unit and by placing the appropriate numerical figure in each geographical unit.

The cartogram shows the principal gas producing regions in india.



CHOICE OF A SUITABLE DIAGRAM

Which diagram out of several ones to select in a given situation is a ticklish problem. The choice would primarily depend upon two factors, namely : (*i*) the nature of the data : and (*ii*) the type of people for whom the diagram is meant. The nature of data would depend whether to use one-dimensional, two-dimensional or three-dimensional diagram, and if it is one-dimensional, whether to adopt the simple bar or sub-divided bar, multiple bar or some other type. As already stated, a cubic diagram would be preferred to a bar if the magnitudes of the figures are very wide apart. The type of people for whom the diagram is intended must also be considered. For example, for drawing attention of an uneducated mass pictograms and cartograms are more effective than cubes, circles, etc. Different types of diagrams such as bars rectangles, cubes, pictograms, pie charts have specific uses. However, bar diagrams are most popular in practice. There are different types of bars and the appropriate type of the bar chart can be decided on the following basis :

(*a*) Simple bar charts should be used where changes in totals are required to be conveyed.

(*b*) Component bar charts are more useful where changes in totals as well as in the size of component figures (absolute ones) are required to be displayed.

(*c*) Percentage sub-divided bar charts are better suited where changes in the relative size of component figures are to be exhibited.

(*d*) Multiple bar charts should be used where changes in the absolute values of the component figures are to be emphasised and the overall total is of no importance.

However, multiple and component bar charts should be used only when there are not more than three or four components, as a large number of components make the bar charts too complex to enable worthwhile visual impression to be gained. When a large number of components have to be shown a pie chart is more suitable.

A pie chart is particularly useful where it is desired to show the relative proportion of the figures that go to make up a single overall total. Unlike bar charts it is not restricted to three or four component figures although its effectiveness tends to dwindle with more than seven or eight components.

Pie charts cannot be used effectively where a series of figures is involved, as a number of different pie charts are not easy to compare. Nor should changes in one overall total be shown by changing the size of the 'pie'.

Occasionally, circles are used to represent size. But it is difficult to compare them and they should not be used when it is possible to use bars. This is because it is easier to compare the lengths of lines or bars than to compare areas or volumes.

Pictograms and cartograms are very elementary forms of visual presentation. However, they are more informative and more effective than other forms for presenting data to the general public who, by and large, neither possess much ability to understand nor take interest in the less attractive forms of presentation. The pictogram is admirably suited to the illustrations of exhibits or articles in newspapers and magazines or for dressing up annual reports. Cartograms or statistical maps are particularly effective in bringing out the geographical pattern that may be concealed in the data.

(iii) Graphs

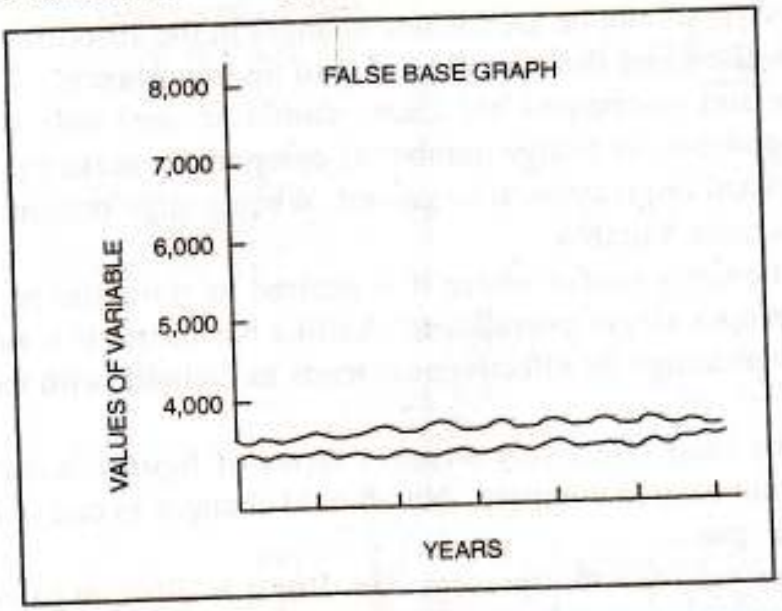
A large variety of graphs are in practical use. However, we shall discuss only some important ones which are more popularly used in practice. Broadly, the various graphs can be divided under the following two heads :

1. Graphs of time series or line graphs.
2. Graphs of frequency distributions.

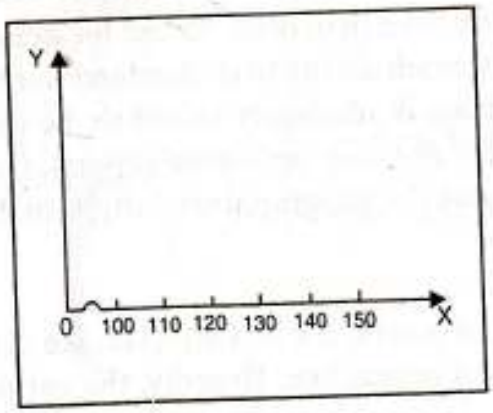
GRAPHS OF TIME SERIES OR LINE GRAPHS

When we observe the values of a variable at different points of time the series so form known as time series. The technique of graphic presentation is extremely helpful in anal changes at different points of time. On the X-axis we generally take the time and on the Y-axi value of the variable and join the various points by straight lines. The graph so formed is know line graph. Such graphs are most widely used in practice. They are simplest to understand, ea to make and most adaptable to many uses. Also several variables can be shown on the same g and a comparison can be made.

One of the fundamental rules while constructing graphs is that the scale on the Y-axis sh begin from zero even if the lowest Y-figures associated with any X-period or value is far a zero. However, if this rule is strictly followed the curve would be very much pulled up towar right, *i.e.*, away from the point of origin. When the gap between zero and smallest value o variable is large, for example, if the variable starts from 50,000, a lot of space would be req to show the variable. It is in order to solve this difficulty that the use of false base is made. Wh a false base has been used, the space between zero and the smallest value of the variable is omitte To bring out this fact clearly that the false base has been used, two zig-zag horizontal lines a drawn above the base line as shown in the graph below :



Just as we have talked of Y-axis starting from zero, we also talk of X-axis starting from ze To represent false base on the X-axis, we draw a kinked line as follows :



It is clear from above that a considerable saving in space is possible in case the variable st from a value much away from 0. It may, however, be noted there is a growing feeling that there

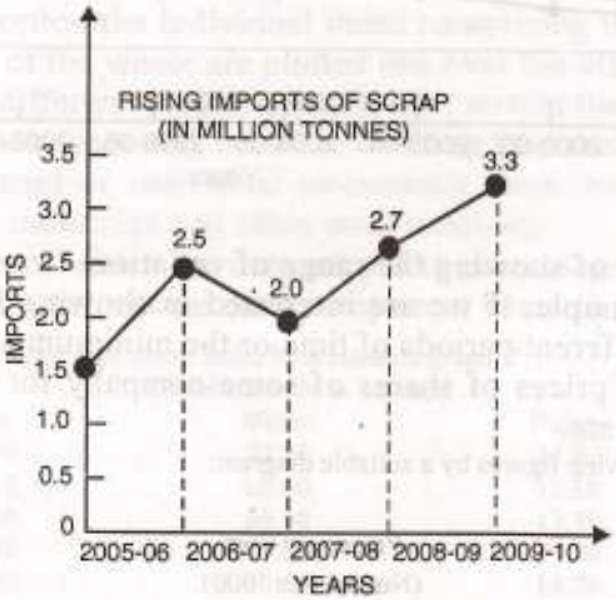
no sanctity in X-axis and Y-axis starting from zero, they can well start from the lowest value or near about from the data given. If that is so the false base line and kinked line need not be used.

Illustration 14. The following data relate to the imports of scrap of Jharkhand Sponge Iron Limited :

Year	:	2005-06	2006-07	2007-08	2008-09	2009-10
Imports (in million tonnes)	:	1.5	2.5	2.0	2.7	3.3

Represent the data graphically.

Solution. The given data can be represented by a graph as shown below :



If the unit of measurement is the same, we can represent two or more variables on the same graph. This facilitates comparison. However, when the number of variables is very large (say, exceeding five or six) and they are all shown on the same graph the chart becomes quite confusing because different lines may cut each other and make it difficult to understand the behaviour of the variables. Therefore, for the sake of clarity we should not represent more than 5 or 6 variables on the same graph. When two or more variables are shown on the same graph, it is desirable to use thick, thin, broken, dotted lines, etc., to distinguish between the various variables.

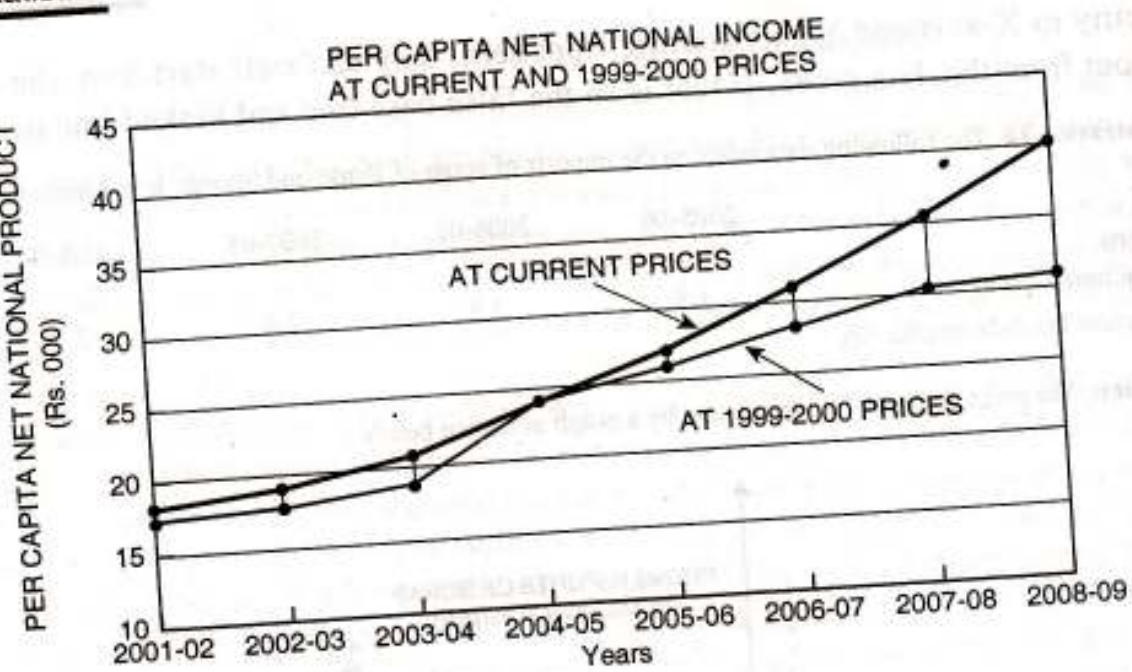
Illustration 15. Represent the following data by a suitable graph.*

PER CAPITA NATIONAL PRODUCT* (Rs.)

Year	*At Current Prices*	*At 1999-2000 Prices*
2001-02	17782	16769
2002-03	18885	17109
2003-04	20871	18301
2004-05	24095	24095
2005-06	27183	25969
2006-07	31080	28074
2007-08	35430	30316
2008-09	40141	31821

*Source : Economic Survey 2009-10, Govt. of India A-3.

Solution.



Range Chart

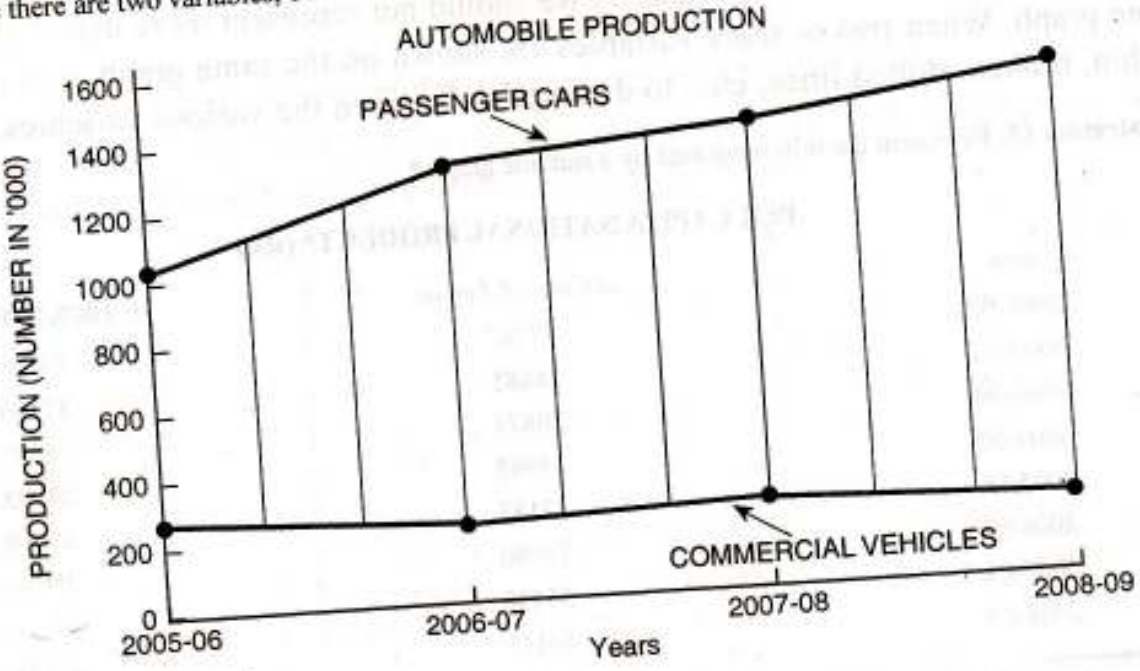
It is a very good method of showing the range of variation, *i.e.*, the minimum and maximum values of a variable. For example, if we are interested in showing the minimum and maximum prices of a commodity for different periods of time or the minimum and maximum temperature or the minimum and maximum prices of shares of some company for different periods, the range chart would be very appropriate.

Illustration 16. Show the following figures by a suitable diagram:

Automobile Production

Year	Passenger Cars (Numbers in '000)	Commercial Vehicles (Numbers in '000)
2005-06	1046	263
2006-07	1323	222
2007-08	1426	246
2008-09	1517	218

Solution. Since there are two variables, both the variables we will show on the same graph.



* *Economic Survey* : Govt. of India, 2008-09 p. 212

The following are the steps in constructing such a chart :

1. Take the time on X-axis and the variable on the Y-axis.

2. Draw two curves by plotting the given data–one curve representing the highest values and the other one lowest values. In the given case, curve A represents lowest prices, whereas curve B highest prices. The gap between curves A and B represents the range of variation.

3. For highlighting the difference between the lowest and highest values the use of some colour or shade, etc., should be made.

Band Graphs

A band graph is a type of line graph which shows the total for successive time periods broken up into sub-totals for each of the component parts of the total. In other words, the band graph shows how and in what proportion the individual items comprising the aggregate are distributed. The various component parts of the whole are plotted one over the other and the gaps between the successive lines are filled by different shades, colours, etc., so that the chart has the appearance of a series of bands. Such a chart is especially useful in dividing total costs into component cost, total sales into department or district or individual salesman's sales, total production by nature of commodity, States, plants, or industries and other such relations.

Band graph can also be used where the data are put to percentage form ; the whole chart will depict 100% and the bands, the percentage each component bears to whole.

Illustration 17. Present the following data about India by a suitable graph :*

(*Production in m. tonnes*)

Year	*Rice*	*Wheat*	*Pulses*	*Coarse cereals*
2003-04	88.50	72.20	14.90	37.60
2004-05	81.10	68.60	13.10	33.50
2005-06	91.80	69.40	13.40	34.10
2006-07	93.40	75.80	14.20	33.90
2007-08	96.69	78.57	14.76	40.76
2008-09	99.15	80.58	14.66	39.48

Solution. The above data can be most suitably presented through a band graph. The procedure of constructing such a graph is as follows :

1. Take the years on the X-axis and the variable on the Y-axis.

2. Plot the various points for different years for rice and join them by straight lines. This is represented by curve *A*.

3. Add the figures of rice for various years to the figures of wheat and plot the points and join them by straight lines. This is represented by curve *B*. The difference between the two curves, *i.e.*, *B* and *A*, gives the production of wheat.

4. Add the figures of rice and wheat to pulses and plot the various points. This is represented by curve *C*. The difference between curve *C* and curve *B* represents production of the pulses.

5. Add the figures of rice, wheat and pulses to other cereals, and draw a curve. This is represented by curve *D*. The difference between curve *D* and curve *C* gives the production figures for other cereals.

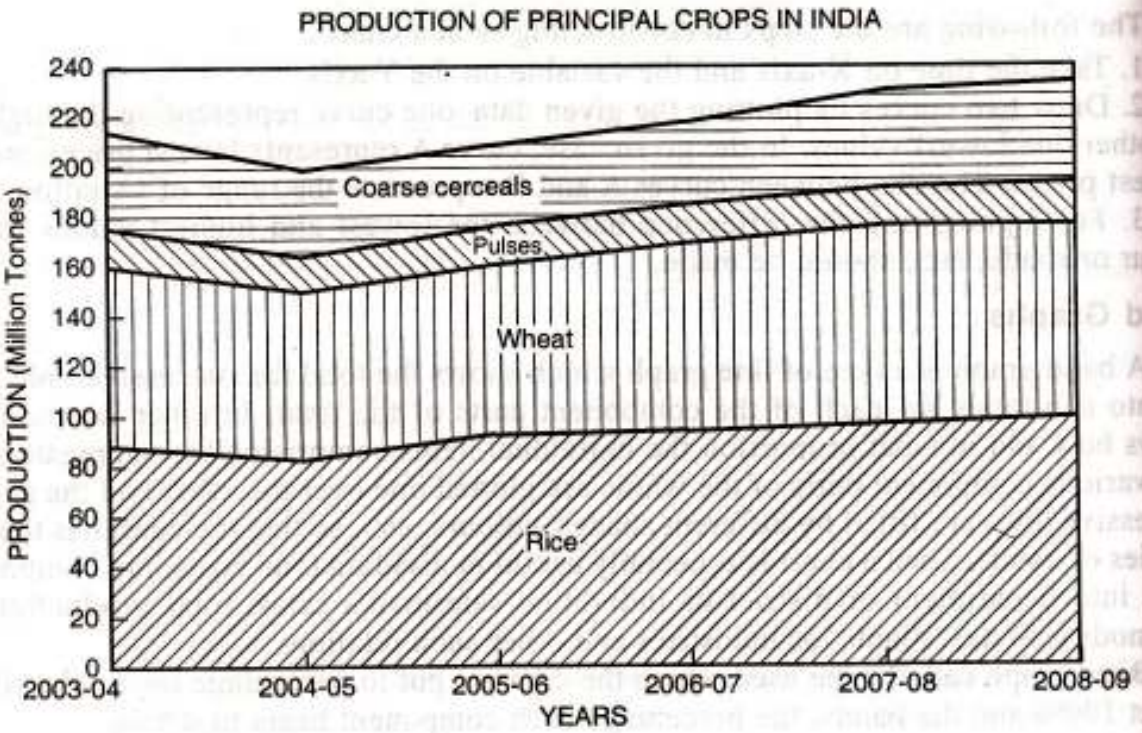
GRAPHS OF FREQUENCY DISTRIBUTIONS

A frequency distribution can be presented graphically in any of the following ways :

1. Histogram.
2. Frequency polygon.
3. Smoothed frequency curve.
4. Cumulative frequency curves or 'Ogives'.

---

* *Economic Survey* : Govt. of India, 2009-10, p. 181.



(1) Histogram

Out of several methods of presenting a frequency distribution graphically, histogram or the column diagram, as it is sometimes called, is the most popular and widely used in practice. The statistical meaning of histogram is that it is a graph that represents the class frequencies in a frequency distribution by vertical adjacent rectangles. A histogram is a graphical method for presenting data, where the observations are located on a horizontal axis (usually grouped into intervals) and the frequency of those observations is depicted along the vertical axis.

While constructing histogram the variable (class interval) is always taken on the X-axis and the frequencies depending on it on the Y-axis. Each class is then represented by a distance of the scale that is proportional to its class-interval. The distance for each rectangle on the X axis shall remain the same in case the class-intervals are uniform throughout ; if they are different the width of the rectangles shall also vary. The Y-axis represents the frequencies of each class which constitute the height of its rectangle. In this manner, we get a series of rectangles each having a class-interval distance as its width and the frequency distance as its height. The area of the histogram represents the total frequency as distributed throughout the classes.

The histogram should be clearly distinguished from a bar diagram. The distinction lies in the fact that whereas a bar diagram is one-dimensional, *i.e.*, only the length of the bar is material and not the width ; a histogram is two-dimensional, that is, in a histogram both the length as well as the width are important.

The histogram is most widely used for graphical presentation of a frequency distribution. However, we cannot construct a histogram for distributions with open-end classes. Moreover, a histogram can be quite misleading if the distribution has unequal class-intervals and suitable adjustments in frequencies are not made.

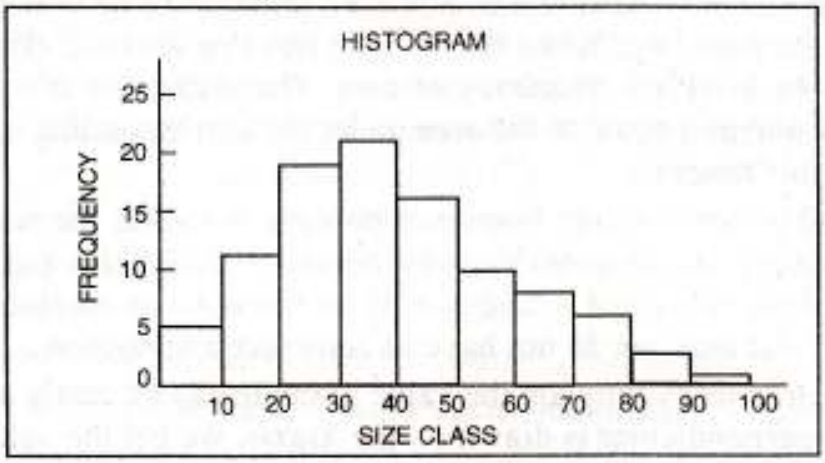
The technique of constructing histogram is now illustrated (*i*) for distributions having equal class-intervals, and (*ii*) for distributions having unequal class-intervals.

Construction of Histogram when Class-Intervals are Equal. When class-intervals are equal, take frequency on the Y-axis, the variable on the X-axis and construct adjacent rectangles. In such a case, the heights of the rectangles will be proportional to the frequencies.

Illustration 18. Represent the following data by a histogram :

Size class	*Frequency*	*Size class*	*Frequency*
0—10	5	50—60	10
10—20	11	60—70	8
20—30	19	70—80	6
30—40	21	80—90	3
40—50	16	90—100	1

Solution. The histogram of the above data is given below.

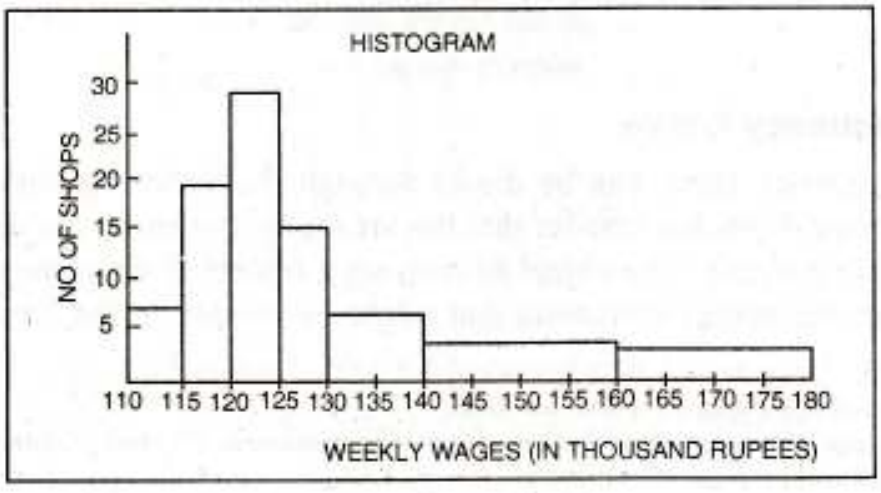


Construction of Histogram when Class-Intervals are Unequal. When class-intervals are unequal, the frequencies must be adjusted before constructing the histogram. For making the adjustment we take that class which has the lowest class-interval and adjust the frequencies of other classes in the following manner. If one class-interval is twice as wide as the one having lowest class-interval, we divide the height of its rectangle by two ; if it is three times more, we divide the height of its rectangles by three, etc., *i.e., the heights will be proportional to the ratios of the frequencies to the width of the classes.* Great care is needed for construction of histograms with unequal class-interval widths. Quite often they are illconstructed and misinterpreted. This would be clear from the following example :

Illustration 19. Represent the following data by means of a histogram :

Weekly profits (in 000's Rs.)	*No. of shops*	*Weekly profits (in 000's Rs.)*	*No. of shops*
110—115	7	130—140	12
115—120	19	140—160	12
120—125	27	160—180	8
125—130	15		

Solution. Since the class-intervals are unequal, frequencies must be adjusted otherwise the histogram would give a misleading picture. The adjustment is done as follows : The lowest class-interval is 5. The frequency of the class 130–140 shall be divided by two since the class-interval is double, that of 140–160 by 4, and so on.



2. Frequency Polygon*

A frequency polygon is a graph of frequency distribution. It has more than four sides. It is particularly effective in comparing two or more frequency distributions. There are two ways in which a frequency polygon may be constructed :

1. We may draw a histogram of the given data and then join by straight lines the mid-points of the upper horizontal side of each rectangle with the adjacent rectangle. The figure so formed is called frequency polygon. Some statisticians, however, prefer to close both the ends of the polygon by extending them to the base line. When this is done two hypothetical classes at each end would have to be included—each with a frequency of zero. This extension is made with the object of making the area under polygon equal to the area under the corresponding histogram. The students are advised to follow this practice.

2. Another method of constructing frequency polygon is to take the mid-points of the various class-intervals and then plot the frequency corresponding to each point and to join all these points by straight lines. The figure obtained would exactly be the same as obtained by the other method. The only difference is that here we do not have to construct a histogram.

By constructing a frequency polygon the value of mode can be easily ascertained. If from the apex of the polygon a perpendicular is drawn on the X-axis, we get the value of mode. Moreover, frequency polygons facilitate comparison of two or more frequency distributions on the same graph.

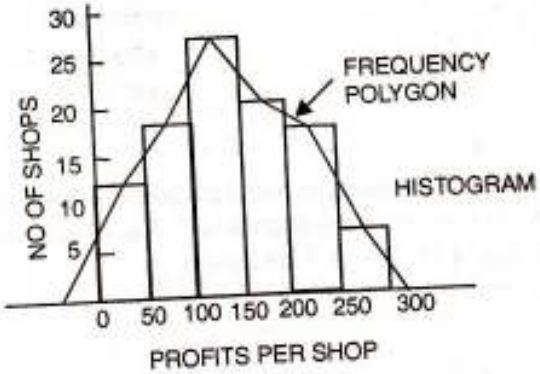
Frequency polygon has a special advantage over the histogram. The frequency polygons of several distributions may be plotted on the same axis, thereby making certain comparisons possible, whereas histograms cannot be usefully employed in the same way. To compare histograms, we must have a separate graph for each. Because of this limitation for purposes in making a graphic comparison of frequency distribution, frequency polygons are preferred.**

Illustration 20. The daily profits (in thousand rupees) of 100 shops are distributed as follows :

Daily Profits :	0—50	50—100	100—150	150—200	200—250	250—300
No. of shops :	12	18	27	20	17	6

Prepare a histogram and frequency polygon of the above data.

Solution. The histogram and frequency polygon of the above data are shown in the diagram below :



3. Smoothed Frequency Curve

A smoothed frequency curve can be drawn through the various points of the polygon. The curve is drawn freehand in such a manner that the area included under the curve is approximately the same as that of the polygon. The object of drawing a smoothed frequency curve is to eliminate as far as possible all accidental variations that might be present in the data. While smoothing

*"Polygon" is a closed figure with more than two sides.

** To make comparison of frequency distributions, percentage frequencies are often preferred. Accordingly, to compare frequency polygons we can plot percentage frequencies.

frequency polygon the fact that it is really derived from the histogram should always be kept in mind. This would imply that the top of the curve would overtop the highest point of the polygon particularly when the magnitude of class-interval is large. The curve should look as regular as possible and all sudden turns should be avoided. The extent of smoothing would, however, depend upon the nature of the data. If it is a natural phenomenon like tossing of coins, smoothing may be freely resorted to as such phenomenon normally has symmetrical curves, but if the phenomenon is social or economic the curve is generally skewed and in such a case smoothing cannot be carried too far.

For drawing smoothed frequency curve it is necessary to first draw the polygon and then smooth it out. As discussed earlier, the polygon can be constructed even without first constructing a histogram by plotting the frequencies at the mid-points of class-intervals. This may save some time but the smoothing of the polygon cannot be done properly without a histogram. Hence, it is desirable to proceed in a sequence, *i.e.*, first to draw a histogram, then a polygon and lastly to smooth it to obtain the smoothed frequency curve. This curve should begin and end at the base line and, as a general rule, it may be extended to the mid-points of the class-intervals just outside the histogram. The area under the curve should represent the total number of frequencies in the entire distribution.

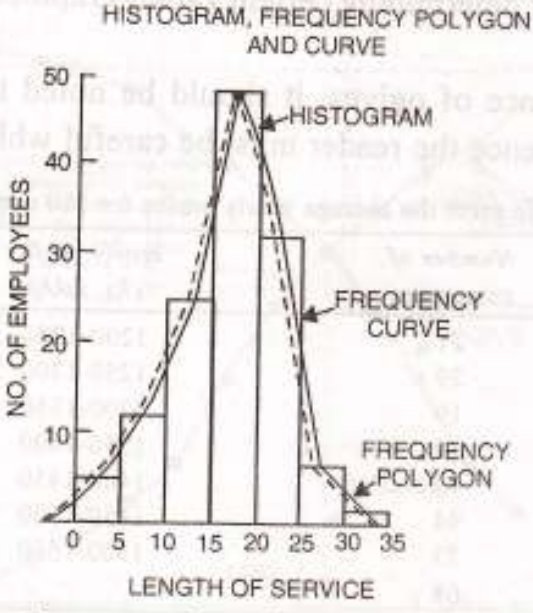
The following points should be kept in mind while smoothing a frequency graph :

1. Only frequency distributions based on samples should be smoothed.
2. Only continuous series should be smoothed.
3. The total area under the curve should be equal to the area under the original histogram or polygon.

Illustration 21. Draw a histogram, frequency polygon and frequency curve representing the following figures :

Length of Service (in years)	*No. of Employees*	*Length of Service (in years)*	*No. of Employees*
Less than 5	5	20—25	32
5—10	12	25—30	6
10—15	25	30—35	1
15—20	48		

Solution. The histogram, frequency polygon and frequency curve of the above data are shown in the following diagram :



When the second method of constructing a frequency polygon is used, the graph would take the same shape as above with a difference that there would be no histogram.

In the construction of frequency polygon, the same difficulties are faced as with histogram, they cannot be used for distributions having open-end classes and suitable adjustment as in case histogram is necessary when there are unequal class-intervals. To get the frequency curve join all points of polygon and smoothen it to obtain the frequency curve as shown in the diagram.

4. Cumulative frequency curves or 'Ogives'

Sometimes one needs to know the answers to questions like 'how many workers of a fact earn more than Rs. 1,500 per month' or 'how many workers earn less than Rs. 1,200 per mont To answer these questions it is necessary to add the frequencies. When frequencies are add they are called cumulative frequencies. These frequencies are then listed in a table called a cumula frequency table. The graph of such a distribution is called a cumulative frequency curve or *Ogive* (pronounced Ogive).

There are two methods of constructing ogive, namely:

(*a*) The 'less than' method, and

(*b*) The 'more than' method.

(*a*) *'Less than' method.* In the 'less than' method, we start with the upper limits of the class and go on adding the frequencies. When these frequencies are plotted, we get a rising curve.

(*b*) *'More than' method.* In the 'more than' method, we start with the lower limits of th classes and from the total frequencies we subtract the frequency of each class. When thes frequencies are plotted, we get a declining curve.

Utility of Ogives. From the standpoint of graphic presentation, the ogive is especially usef for the following purposes:

1. To determine as well as to portray the number or proportion of cases above or below a given value.

2. To compare two or more frequency distributions. Generally there is less overlapping when comparing several ogives on the same grid than when comparing several simple frequency curves in this manner.

3. Ogives are also drawn for determining certain values graphically, such as median, quartiles deciles, etc.

Despite the great significance of ogives, it should be noted that they are not as simple t interpret as one may feel and hence the reader must be careful while using them.

Illustration 22. The following table gives the average yearly profits for 580 companies :

Yearly profits (Rs. lakh)	*Number of companies*	*Yearly profits (Rs. lakh)*	*Number of companies*
800-850	21	1200-1250	36
850-900	29	1250-1300	45
900-950	19	1300-1350	27
950-1000	39	1350-1400	48
1000-1050	43	1400-1450	21
1050-1100	94	1450-1500	12
1100-1150	73	1500-1550	5
1150-1200	68		

(*a*) Draw ogives by 'less than' and 'more than' method for the data given above.

(*b*) Compute Q_1, D_4 and P_{65} and verify your result with the help of the drawn ogive

(c) Find the number of companies whose yearly profits (in Rs. lakh) lie between Rs. 1180 and Rs. 1480

(MBA, GGSIP Univ., 2000, MBA, DU, 2002, 2007)

Solution. Let us first arrange the frequency distribution for 'less than' and 'more than' methods as given below :

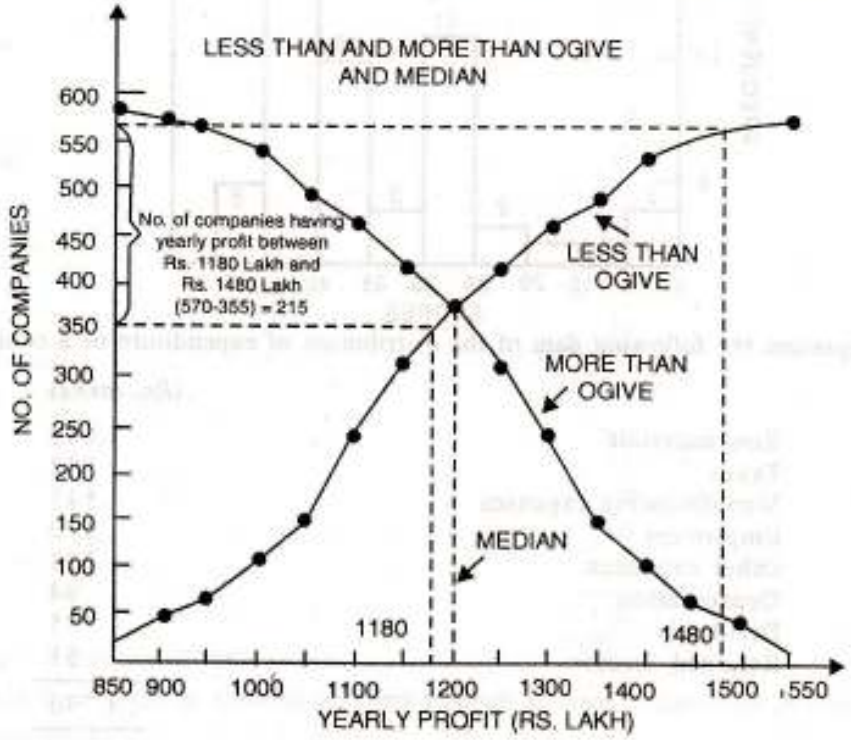
OGIVE BY LESS THAN METHOD

Yearly profits (Rs. lakh)	*Number of companies*	*Yearly profits (Rs. lakh)*	*Number of companies*
Less than 850	21	Less than 1250	422
" " 900	50	" " 1300	467
" " 950	69	" " 1350	494
" " 1000	108	" " 1400	542
" " 1050	151	" " 1450	563
" " 1100	245	" " 1500	575
" " 1150	318	" " 1550	580
" " 1200	386		

OGIVE BY MORE THAN METHOD

Yearly profits (Rs. lakh)	*Number of companies*	*Yearly profits (Rs. lakh)*	*Number of companies*
More than 800	580	More than 1200	194
" " 850	559	" " 1250	158
" " 900	530	" " 1300	113
" " 950	511	" " 1350	86
" " 1000	472	" " 1400	38
" " 1050	429	" " 1450	17
" " 1100	335	" " 1500	5
" " 1150	262		

With the help of these frequency distribution tables, we can draw ogives by less than and more than method as shown below :



It is clear from the above graph that there are 570 companies who are earning profits less than Rs. 1480 lakhs and there are 355 companies who are getting wages less than Rs. 1180 lakhs. Thus the number of companies earning profits between Rs. 1180 lakhs and Rs. 1480 lakhs is (570–355) = 215.

MISCELLANEOUS ILLUSTRATIONS

Illustration 23. You have conducted a market survey with a sample of size 50 regarding the acceptability of a new product which your company wants to launch. The scores of the respondents on the appropriate scale are as follows:

40	45	41	45	45	30	39	8	48	25
26	9	23	24	26	29	8	40	41	42
39	35	18	25	35	40	42	43	44	36
27	32	28	27	25	26	38	37	36	35
32	28	40	41	43	44	45	40	39	41

Prepare a frequency table and present the same information as a histogram. *(MBA, HPU, 2007)*

Solution. Frequency table of the given data is made as follows :

FREQUENCY DISTRIBUTION OF SCORES

Scores	*Tally*	*Frequency*
5-10	III	3
10-15	—	—
15-20	I	1
20-25	II	2
25-30	𝍸 𝍸 I	11
30-35	III	3
35-40	𝍸 𝍸	10
40-45	𝍸 𝍸 𝍸	15
45-50	𝍸	5
		Total 50

HISTOGRAM OF THE GIVEN DATA

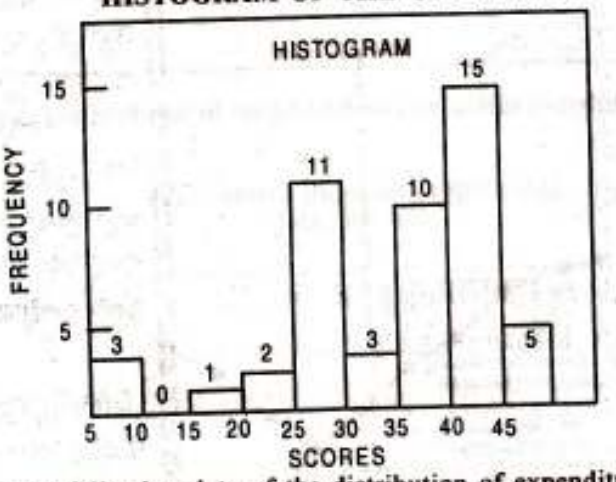


Illustration 24. Represent the following data of the distribution of expenditure of a company by suitable diagram:

	(Rs. lakhs)
Raw materials	1,689
Taxes	582
Manufacturing expenses	543
Employees	470
Other expenses	286
Depreciation	94
Dividends	75
Retained income	51
	3,790

(MBA, KU., 2005)

Solution. The above data can be represented by a pie diagram. Convert the given data in terms of percentage as follows:

Distribution	*Rs. lakhs*	*Percentage*
Raw materials	1,689	44.6
Taxes	582	15.4
Manufacturing expenses	543	14.3
Employees	470	12.4
Other expenses	286	7.5
Depreciation	94	2.5
Dividends	75	2.0
Retained income	51	1.3
	Total 3,790	100.0

PIE DIAGRAM SHOWING THE DISTRIBUTION OF EXPENDITURE

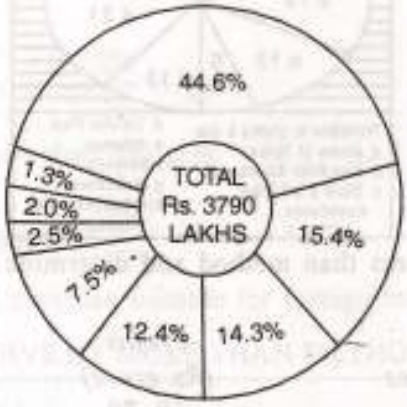


Illustration 25. The following data relate to the Central Budget 2007-08 How a Rupee comes from :

		(in paise)
(a)	Internal borrowings	20
(b)	External assistance	3
(c)	Other capital receipts	6
(d)	Borrowings from RBI	6
(e)	Customs	20
(f)	Excise	21
(g)	Corporation tax	5
(h)	Income tax	5
(i)	Other taxes	2
(j)	Non-tax revenue	12
		100

Where a Rupee goes to?

		(in paise)
(a)	Share of taxes	13
(b)	Non-plan assistance	7
(c)	State & U.T. plan assistance	11
(d)	Central Plan	15
(e)	Interest	21
(f)	Defence	13
(g)	Subsidies	8
(h)	Other non-plan expenditure	12
		100

Represent the data by a suitable diagram.

Solution. Since we are given various subdivisions from which a rupee comes from and is spent, suitable diagram would be a subdivided pie diagram.

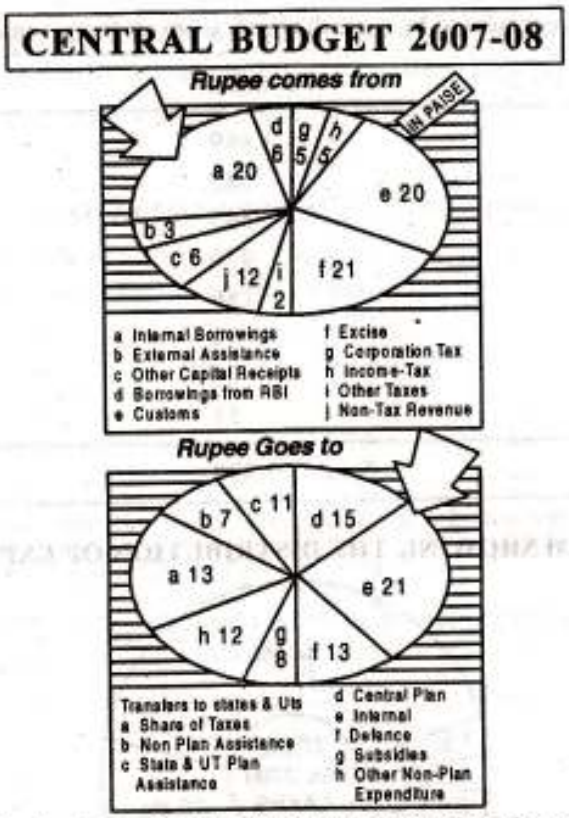


Illustration 26. Draw an ogive by less than method and determine the number of companies getting profits between Rs. 45 crores and Rs. 75 crores:

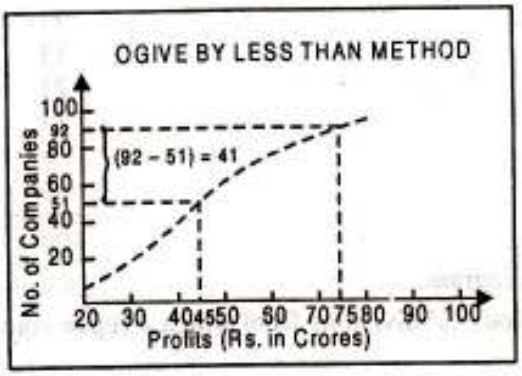
Profits (Rs. crore)	*No. of Companies*	*Profits (Rs. crore)*	*No. of Companies*
10–20	8	60–70	10
20–30	12	70–80	7
30–40	20	80–90	3
40–50	24	90–100	1
50–60	15		

(MBA, DU, 1999)

Solution.

OGIVE BY LESS THAN METHOD

Profits (Rs. crore)	*No. of Companies*
Less than 20	8
" " 30	20
" " 40	40
" " 50	64
" " 60	79
" " 70	89
" " 80	96
" " 90	99
" " 100	100



It is clear from the graph that the number of companies getting profits less than Rs. 75 crores is 92 and the number of companies getting profits less than Rs. 45 crores is 51. Hence the number of companies getting profits between Rs. 45 crores and Rs. 75 crores is 92–51=41.

Illustration 27. The following distribution is with regard to weight in grams of mangoes of a given variety. If mangoes of weight less than 443 grams be considered unsuitable for foreign market, what is the percentage of total yield suitable for it? Assume the given frequency distribution to be typical of the variety:

Weight in gms	*No. of mangoes*	*Weight in gms*	*No. of mangoes*
410–419	10	450–459	45
420–429	20	460–469	18
430–439	42	470–479	7
440–449	54		

Draw an ogive of 'more than' type of the above data and deduce how many mangoes will be more than 443 grams.
(M.B.A., Delhi Univ., 2006)

Solution. Mangoes weighing more than 443 gms. are suitable for foreign market. Number of mangoes weighing more than 443 gms. lies in the last four classes. Number of mangoes weighing between 444 grams and 449 grams would be:

$$\frac{6}{10} \times 54 = \frac{324}{10} = 32.4$$

Total number of mangoes weighing more than 443 gms. = 32.4 + 45 + 18 + 7 = 102.4

$$\text{Precentage} = \frac{102.4}{196} \times 100 = 52.25$$

Therefore, the percentage of the total mangoes suitable for foreign market is 52.25.

OGIVE BY MORE THAN METHOD

Weight more than (gms.)	*No. of mangoes*
410	196
420	186
430	166
440	124
450	70
460	25
470	7

From the graph, it can be seen that there are 103 mangoes whose weight will be more than 443 gms. and are suitable for foreign market.

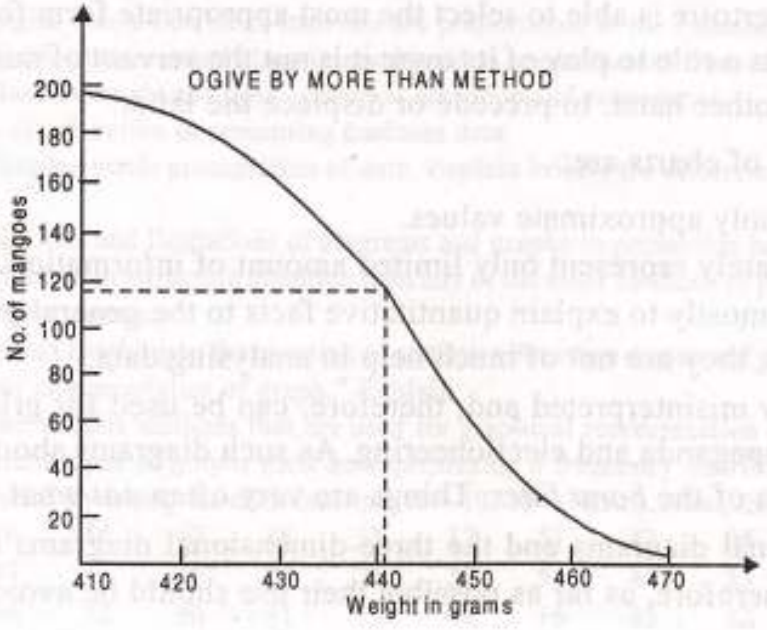
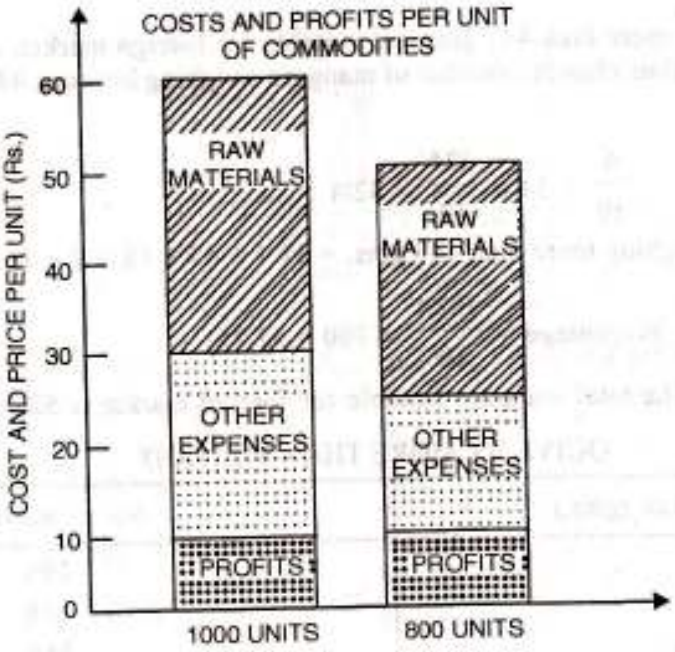


Illustration 28. Draw a rectangular diagram to represent the following information:

	Factory A (Rs.)	*Factory B (Rs.)*
Price per unit	60	50
Units produced	1,000	800
Value of Raw Materials used	3,000	2,400
Other expenses	2,000	1,400
Profit	1,000	1,000

(MBA, HPU, 2006)

Solution.



Limitations of Charts

Although charts are powerful and effective method for presenting statistical data, they are not under all circumstances and for all purposes complete substitutes for tabular and other forms of presentation. The well-trained specialist in the field is one who recognises not only the advantages but also the limitations of these techniques. He knows when to use and when not to use these methods and from his repertoire is able to select the most appropriate form for every purpose. It is said "Graphic statistics has a role to play of its own; it is not the servant of numerical statistics, but it cannot pretend, on the other hand, to precede or displace the latter."

The main limitations of charts are:

1. They can present only approximate values.
2. They can appropriately represent only limited amount of information.
3. They are intended mostly to explain quantitative facts to the general public. From the point of view of the statistician, they are not of much help in analysing data.
4. They can be easily misinterpreted and, therefore, can be used for grinding one's own axe during advertisement, propaganda and electioneering. As such diagrams should never be accepted without a close inspection of the *bona fides*. Things are very often *not* what they appear to be.
5. The two-dimensional diagrams and the three-dimensional diagrams cannot be accurately appraised visually and, therefore, as far as possible their use should be avoided.

PROBLEMS

Answer the following questions, each question carries **one** mark.

(*i*) State the use of pie diagram.

(*ii*) What is the use of histogram ? (*MBA, Madurai Kamraj, Nov. 2001*)

(*iii*) What do you mean by a pictogram ? (*MBA, Madurai Kamraj, Nov. 2003*)

(*iv*) Explain how the ogives are drawn for any frequency distribution.

(*v*) What is tabulation ?

(*vi*) What is a percentile ?

(*vii*) What is cumulative frequency curve ?

(*viii*) What is open end distribution ?

(*ix*) What is histogram ?

(*x*) What is frequency polygon ?

Answer the following questions, each question carries **four** marks. (*MBA, UP Tech. Univ, 2004*)

(*i*) Distinguish between classification and tabulation of data.

(*ii*) Give at least two uses each of classification and tabulation.

(*iii*) Distinguish between discrete and continuous variables with suitable examples.

(*iv*) Distinguish between one dimensional and two dimensional diagrams.

(*v*) What are the important steps in forming a frequency distribution ?

(*a*) Explain the role of tabulation in presenting business data, and discuss briefly the different methods of presentation.

(*b*) What are the different types of graphs and charts known to you? What are their uses?

(*a*) Explain the term 'classification' and 'tabulation'. Point out their importance in a statistical investigation. What precautions would you take in tabulating statistical data?

(*b*) "In classification and tabulation, commonsense is the chief requisite and experience the chief teacher." Comment.

(*c*) Explain pie diagram and histogram as methods of diagrammatic and graphic presentation of data with suitable examples.

(*a*) What are the requisites of a good table? State the rules that serve as a guide in tabulating statistical material.

(*b*) Explain briefly some of the uses of graphs and charts in presenting business data etc. (*MBA, Osmania Univ., 2004*)

(*a*) What are the chief functions of tabulation? What precautions would you take in tabulating statistical data?

(*b*) What are the characteristics of a good table?

(*c*) Explain Struge's rule in forming frequency distribution.

(*a*) Distinguish clearly between a continuous variable and a discrete variable. Give two examples of continuous variables and two examples of discontinuous or discrete variables that might be used by a business statistician.

(*b*) Explain how tables, graphs and charts help in the effective presentation of data.

(*c*) State whether the following statement is *true* or *false* :

The heights of rectangles erected on class-intervals are proportional to the cumulative frequency of the classes.

(*a*) What are the objections to unequal class-intervals and to open classes? State the conditions under which the use of unequal class-intervals and open classes is desirable and necessary.

(*b*) Mention the role of tabulation in presenting business data.

Point out the role of diagrammatic presentation of data. Explain briefly the different types of bar diagrams known to you.

(*a*) Explain clearly the role and limitations of diagrams and graphs in presenting business data.

(*b*) Charts are more effective in attracting attention than any of the other methods of presenting data. Do you agree? Give reasons for your answer.

(*a*) "Diagrams do not add anything to the meaning of statistics but when drawn and studied intelligently they bring to view the salient characteristics of graph." Explain.

(*b*) Explain briefly the various methods that are used for graphical representation of frequency distributions.

(*c*) Explain the different types of graphs used for representing a frequency distribution. (*MBA, KU., 2006*)

For a frequency distribution by taking the class-intervals 15—19, 20—24,.....etc., for the following data:

30	42	30	54	40	48	15	17	51	42	25
41	30	27	42	36	28	26	37	54	44	31
36	40	36	22	30	31	19	48	16	42	32
21	22	40	33	41	21					

12. Form a frequency distribution taking a suitable class-interval for the following data giving the age of 52 employees in a government agency.

67	34	36	48	49	31	61	34	43	45	38
32	27	61	29	47	36	50	46	30	46	32
30	33	45	49	48	41	53	36	37	37	47
30	46	50	28	35	35	38	36	46	43	34
62	69	50	28	44	43	60	39			

13. Draft a blank table to show :

(*a*) Sex, (*b*) three ranks—supervisors, assistants and clerks, (*c*) years 2000 and 2010, and (*d*) Age groups—18 years and under, over 18 years but less than 55 years, over 55 years.

14. Prepare a table with a proper title, division and sub-divisions to represent the following heads of information:

(*i*) Export of Cotton piece-goods from India.
(*ii*) To Burma, China, Indonesia, Iran, Iraq.
(*iii*) Amount of piece-goods to each country.
(*iv*) Value of piece-goods to each country.
(*v*) From 2007-08 to 2009-10, year by year.
(*vi*) Total amount of exports each year.
(*vii*) Total value of exports each year.

15. Represent the following data by a suitable diagram :

Year	2005	2006	2007	2008	2009	2010
Sale of steel (in thousand tonnes)	8	8.8	9.2	10.2	7.6	12.5

16. Represent the following data by a suitable diagram:

Year	2004-2005	2005-06	2006-07	2007-08	2008-09	2009-10
Profit before taxes (Rs. lakh)	28	29.4	30.2	27	32.5	40.6

17. Represent the following data by a sub-division bar diagram :

(in lakh for Rs.)

Year	2007-08	2008-09	2009-10
Gross Income	460	482	552
Gross expenditure	400	450	500
Net Income	60	32	22

18. Represent the following information diagrammatically :

Factory	*Wages* (*Rs.*)	*Material* (*Rs.*)	*Other costs* (*Rs.*)	*Profits* (*Rs.*)	*No. of Units*
A	3,000	5,000	1,000	1,000	1,000
B	2,000	3,000	800	500	700

19. Represent the following data by a suitable diagram :

UTILIZATION OF 100 PAISE OF INCOME BY XYZ LTD. IN THE YEAR 2009-10

1.	Raw Material, Manufacturing and other Expenses	40 Paise
2.	Wages, Salaries, Bonus and other Benefits to employees	12 Paise
3.	Selling and Distribution Expenses	4 Paise
4.	Interest–Financing Charges	4 Paise
5.	Depreciation and Development Rebate	3 Paise
6.	Excise Duty of Sales	15 Paise
7.	Taxation	13 Paise
8.	Dividends	6 Paise
9.	Surplus retained in Business	3 Paise

20. Represent the following data by a "Pie Diagram" :
CHEQUES CLEARED IN INDIA IN CLEARING HOUSES IN THE YEARS 2009 AND 2010

Centres	*Amount in Crore of rupees*	
	2009	*2010*
Mumbai	829	2,670
Kolkata	1,070	2,443
Chennai	108	274
Other centres	313	615
Total	2,320	6,002

21. Draw a suitable diagram for the following :

Expenditure	*Family A (Income Rs. 10000)*	*Family B (Income Rs. 12000)*
Food	3000	4000
Clothing	2500	2000
Education	500	3600
Others	3800	3000
Saving or deficit	+200	−600

22. Draw the histogram, frequency curve and the ogive curve for the following data pertaining to income distribution for 1500 employees working in a company.

Monthly income (in thousand Rs.)	*No. of employees*	*Monthly income (in thousand Rs.)*	*No. of Employees*
18–20	10	28–30	320
20–22	35	30–32	200
22–24	140	32–34	75
24–26	300	34–36	35
26–28	370	36–38	15

23. What is meant by a histogram? State briefly how it is constructed? Indicate clearly how the histogram in respect of the following data can be drawn (only a rough sketch is required). State also how you can draw histogram in respect of unequal class-intervals.

Mid-Value	*Frequency*	*Mid-value*	*Frequency*
115	6	165	60
125	25	175	38
135	48	185	22
145	72	195	3
155	116		

24. The following table gives the total units produced at the beginning of different years. Represent the data graphically and estimate the mid-year value for 2002 and 2010.

Years	*Units Produced*	*Years*	*Units Produced*
2002	20	2007	811
2003	62	2008	1,104
2004	147	2009	1,425
2005	300	2010	1,755
2006	536		

25. Represent the data showing the number of companies in various ranges of subscribed capital by means of a histogram:

Subscribed capital (Rs. crore)	*No. of companies*	*Subscribed capital (Rs. crore)*	*No. of companies*
Up to 10	10	50–80	7
10–20	12	80–100	8
20–30	10	Above 100	5
30–50	14		

26. The data below give the yearly profits (in thousand rupees) of two companies *A* and *B*:

	Profits (In '000 Rs.)	
Year	*Company A*	*Company B*
2005–06	120	90
2006–07	135	95
2007–08	140	108
2008–09	160	120
2009–10	175	130

Represent the data by means of a suitable diagram.

27. Below is given the frequency distribution of weekly wages of 100 workers in a factory:

Monthly wages (Rs.)	*No. of workers*	*Monthly wages (Rs.)*	*No. of workers*
3000–3500	3	5500–6000	10
3500–4000	5	6000–6500	8
4000–4500	12	6500–7000	5
4500–5000	23	7000–7500	3
5000–5500	31		

Draw the ogive for the distribution and use it to determine the median wage of a worker.

28. Present the following information in a suitable tabular form supplying the figures not directly given:

"In 2009, out of a total 2,000 workers in a factory 1,550 were members of a trade union. The number of women workers employed was 250, out of which 200 did not belong to any trade union."

"In 2010, the number of union workers was 1,725 of which 1,600 were men. The number of non-union workers was 380 among which 155 were women." *(MBA, GGSIP Univ.)*

29. Draw an Ogive for the following distribution. Read the median from the graph and verify the result by formula. How many workers earned monthly wages between Rs. 5,400 and Rs. 5470 ?

Monthly wages (in Rs.)	*No. of workers*	*Monthly wages (in Rs.)*	*No. of workers*
5000–5200	6	5800–6000	16
5200–5400	10	6000–6200	12
5400–5600	22	6200–6400	15
5600–5800	30		

30. The proprietor of Goodwill Tyres Co. kept a record of the number of car tyres of each brand that were sold during 2009–10. He arranged the data as follows :

Brand	*No. of Tyres Sold*
Dunlop	280
Modi	270
Firestone	200
Ceat	190
Goodyear	160
J.K.	100

(*a*) What kind of a distribution is this?

(*b*) What are the class boundaries of each class?

(*c*) Present the data by a suitable diagram/graph.

31. Draw a suitable diagram to represent the following information :

	Company A	*Company B*
Selling price	12,000	8,000
Raw Materials	5,000	6,000
Direct Wages	4,000	3,200
Factory and office on cost	1,000	800

32. A company dealing in 60 products, in the course of establishing an inventory control system, classified products according to price as shown in the frequency table below :

Unit cost (in hundreds of Rs.) :	3–5	6–8	9–11	12–14	15–17	18–20	21–23
No. of items :	6	8	10	20	8	5	3

Prepare more than ogive for the distribution on a graph paper. Use this graph to determine 20th and 80th percentiles.

33. Discuss the following terms with illustrations :
   (*i*) Classification and Tabulation.
   (*ii*) Frequency, cumulative frequency and frequency polygon.
   (*iii*) Histogram, line and bar diagrams.
   (*iv*) Pie chart. *(MBA, Vikram Univ., 1998)*

34. (*a*) Diagrams help us to visualise the whole meaning of a numerical complex at a single glance". Comment.

   (*b*) Draw a suitable diagram to represent the following:

	Selling Price Per Unit (Rs.)	*Qty Sold*	*Wages*	*Materials*	*Others*
Factory A	400	20	3200	2400	1600
Factory B	600	30	6000	6000	9000

Show also profit or loss as the case may be. *(MBA, HPU, 2001)*

35. From the following frequency distribution, prepare the less than and more than cumulative frequency curve (ogive curve)

Class-Interval :	0–10	10–20	20–30	30–40	40–50	50–60
Frequency :	8	12	30	25	18	17

(MBA, K.U., 2003)

*****

4

Measures of Central Tendency

In the previous chapters, data collection and presentation of data were discussed. Even after the data have been classified and tabulated one often finds too much details for many uses that may be made of the information available. We, therefore, frequently need further analysis of the tabulated data. One of the powerful tools of analysis is to calculate a single *average value* that represents the entire mass of data. The word average is very commonly used in day-to-day conversation. For example, we often talk of average work, average income, average age of employees, etc. An 'average' thus is a single value which is considered as the most representative or typical value for a given set of data. Such a value is neither the smallest nor the largest value, but is a number whose value is somewhere in the middle of the group. For this reason an average is frequently referred to as a measure of central tendency or central value. Measures of central tendency show the tendency of some central value around which data tends to cluster.

OBJECTIVES OF AVERAGING

There are two main objectives of the study of averages :

(*i*) *To get one single value that describes the characteristics of the entire data.* Measures of central value, by condensing the mass of data in one single value, enable us to get an idea of the entire data. Thus one value can represent thousands, lakhs and even millions of values. For example, it is impossible to remember the individual incomes of millions of earning people of India and even if one could do it there is hardly any use. But if the average income is obtained, we get one single value that represents the entire population. Such a figure would throw light on the standard of living of an average Indian.

(*ii*) *To facilitate comparison.* Measures of central value, by reducing the mass of data in one single figure, enable comparisons to be made. Comparison can be made either at a point of time or over a period of time. For example, the figure of average sales for December may be compared with the sales figures of previous months or with the sales figure of another competitive firm.

CHARACTERISTICS OF A GOOD AVERAGE

Since an average is a single value representing a group of values, it is desirable that such a value satisfies the following properties :

(*i*) *It should be easy to understand.* Since statistical methods are designed to simplify complexity it is desirable that an average be such that can be readily understood, its use is bound to be very limited.

(*ii*) *It should be simple to compute.* Not only an average should be easy to understand but also should be simple to compute so that it can be used widely. However, though ease of computation is desirable, it should not be sought at the expense of other advantages, *i.e.*, if in the interest of greater accuracy, use of a more difficult average is desirable one should prefer that.

(*iii*) *It should be based on all the observations.* The average should depend upon each and every observation so that if any of the observation is dropped average itself is altered.

(*iv*) *It should be rigidly defined.* An average should be properly defined so that it has one and only one interpretation. It should preferably be defined by an algebraic formula so that if different people compute the average from the same figures they all get the same answer (barring arithmetical mistakes).

(*v*) *It should be capable of further algebraic treatment.* We should prefer to have an average that could be used for further statistical computations. For example, if we are given separately the figures of average income and number of employees of two of more companies we should be able to compute the combined average.

(*vi*) *It should have sampling stability.* We should prefer to get a value which has what the statisticians call 'sampling stability'. This means that if we pick 10 different group of college students, and compute the average of each group, we should expect to get approximately the same values. It does not mean, however, that there can be no difference in the value of different samples. There may be some difference but those averages in which this difference, technically called 'sampling fluctuation,' is less are considered better than those in which this difference is more.

(*vii*) *It should not be unduly affected by the presence of extreme values.* Although each and every observation should influence the value of the average, none of the observations should influence it unduly. If one or two very small or very large observations unduly affect the average, *i.e.*, either increase its value or reduce its value, the average cannot be really typical of the entire set of data. In other words, extremes may distort the average and reduce its usefulness.

The following are the important measures of central tendency which are generally used in business :

A. Arithmetic mean, B. Median, C. Mode, D. Geometric mean, and E. Harmonic mean

A. ARITHMETIC MEAN

The most popular and widely used measure for representing the entire data by one value is what most laymen call an 'average' and what the statisticians call the arithmetic mean. Its value is obtained by adding together all the observations and by dividing this total by the number of observations.

Calculation of Arithmetic Mean—Ungrouped Data

For ungrouped data, arithmetic mean may be computed by applying any of the following methods :

(*i*) Direct method,

(*ii*) Short-cut method.

(i) Direct Method : The arithmetic mean, often simply referred to as mean, is the total of the values of a set of observations divided by their total number of observations. Thus, if $X_1, X_2, \ldots\ldots, X_N$ represent the values of N items or observations, the arithmetic mean denoted by $\bar{X}$ is defined as :

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

If the subscripts are dropped, the formula becomes :

$$\bar{X} = \frac{\Sigma X}{N}$$

It may be pointed out that in keeping with standard statistical practice, the symbol $\bar{X}$ will represent throughout this text the arithmetic mean of a set of observations.

Illustration 1. The monthly income (in rupees) of 10 employees working in a firm is as follows :

4487 4493 4502 4446 4475 4492 4572 4516 4468 4489

Find the average monthly income.

Solution. Let income be denoted by X.

$\Sigma X = 4487+4493+4502+4446+4475+4492+4572+4516+4468+4489 = 44,940$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{44940}{10} = 4494$$

Hence the average monthly income is Rs. 4,494.

(ii) Short-cut Method. The arithmetic mean* can also be calculated by taking deviations from any arbitrary point in which case the formula shall be :

$$\bar{X} = A + \frac{\Sigma d}{N}$$

where $d = (X - A)$

and A = Arbitrary point (or assumed mean).

It should be noted that any value can be taken as arbitrary point and the answer would be the same as obtained by the direct method.

Illustration 2. Calculate average monthly income by the short-cut method from data of Illustration 1 taking deviations from 4460 as the arbitrary point.

Solution.

CALCULATION OF AVERAGE INCOME

X (Rs.)	$(X-4460)$ d
4487	+27
4493	+33
4502	+42
4446	−14
4475	+15
4492	+32
4572	+112
4516	+56
4468	+ 8
4489	+29
	$\Sigma d = +340$

$$\bar{X} = A + \frac{\Sigma d}{N} = 4460 + \frac{340}{10} = 4460 + 34 = \text{Rs. } 4494.$$

One may find that the short-cut method takes more time as compared to direct method. However, this is true only for ungrouped data. In case of grouped data, considerable saving in time is possible by adopting the short-cut method.

Calculation of Arithmetic Mean—Grouped Data

For grouped data, arithmetic mean may be computed by applying any of the following methods :

(i) Direct method,

(ii) Short-cut method.

---

*This formula is derived as follows :

Let $d = X - A$ or $X = A + d$

Taking summation of both sides and dividing by N, we get

$$\frac{\Sigma X}{N} = \frac{\Sigma A}{N} + \frac{\Sigma d}{N} \quad \text{or} \quad \bar{X} = A + \frac{\Sigma d}{N}$$

The population mean is denoted by μ (μ is the Greek letter mu) and the sample mean by $\bar{X}$.

(*i*) **Direct Method.** When direct method is used

$$\bar{X} = \frac{\sum fX}{N}$$

where X = mid-point of various classes.
f = the frequency of each class.
N = the total frequency.

Note. For computing mean in the case of grouped data the mid-points of the various classes are taken as representative of that particular class. The reason is that when the data are grouped, the exact frequency with which each value of the variable occurs in the distribution in unknown. We only know the limits within which a certain number of frequencies occur. For example, when we say that the number of persons within the income group 4000–4500 is 50 we cannot say as to how many persons out of 50 are getting 4001, 4002, 4003, etc. We, therefore, make an assumption while calculating arithmetic mean that the frequencies within each class are distributed uniformly or evenly over the range of the class-interval, *i.e.*, there will be as many observations below the mid-point as above it. Unless such an assumption is made, the value of mean cannot be computed.

Illustration 3. The following are the figures of profits earned by 1,400 companies during 2003-04.

Profits (Rs. Lakhs)	*No. of Companies*	*Profits (Rs. Lakhs)*	*No. of Companies*
200–400	500	1,000–1,200	100
400–600	300	1,200–1,400	80
600–800	280	1,400–1,600	20
800–1000	120		

Calculate the average profits for all the companies.

Solution.

CALCULATION OF AVERAGE PROFITS

Profits (Rs. Lakhs)	*Mid-points X*	*No. of Companies f*	*fX*
200–400	300	500	1,50,000
400–600	500	300	1,50,000
600–800	700	280	1,96,000
800–1,000	900	120	1,08,000
1,000–1,200	1100	100	1,10,000
1,200–1,400	1300	80	1,04,000
1,400–1,600	1500	20	30,000
		$N = 1,400$	$\Sigma fX = 8,48,000$

$$\bar{X} = \frac{\sum fX}{N} = \frac{8,48,000}{1,400} = 605.71$$

Thus, the average profit is Rs. 605.71 lakhs.

• $\bar{X} = \frac{\sum fX}{N}$ (direct method)

Now $d = \frac{X - A}{i}$, $\therefore \quad X = A + id$

Substituting the value of X in the direct method, we get

$$\bar{X} = \frac{\sum f(A+id)}{N} = \frac{\sum fA}{N} + \frac{\sum fd}{N} \times i$$

Hence $$\bar{X} = A + \frac{\sum fd}{N} \times i \qquad (\because \Sigma f = N)$$

(*ii*) **Short-cut Method*.** When short-cut method is used, the following formula is applied :

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i$$

where $$d = \frac{X - A}{i}$$

and i = size of the equal class interval.

Illustration 4. Calculate the average profit by the short-cut method from the data of Illustration 3.

Solution. **CALCULATION OF AVERAGE PROFITS**

Profits (Rs. Lakhs)	*Mid points X*	*f*	*(X–900)/200 d*	*fd*
200–400	300	500	– 3	– 1,500
400–600	500	300	– 2	– 600
600–800	700	280	– 1	– 280
800–1000	**900**	120	0	0
1000–1200	1100	100	+ 1	+ 100
1200–1400	1300	80	+ 2	+ 160
1400–1600	1500	20	+ 3	+ 60
		N = 1,400		Σfd = – 2,060

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 900 - \frac{2{,}060}{1{,}400} \times 200 = 900 - 294.29 = \text{Rs. } 605.71$$

Hence the average profit is Rs. 605.71 lakhs.

Correcting Incorrect Values

It sometimes happens that due to an oversight or mistake in copying certain wrong values are taken while calculating the mean. The problem is how to find out the correct mean. The process is very simple. From ΣX deduct wrong observations and add correct observations and then divide the correct ΣX by the number of observations and the result so obtained will give the value of the correct mean.

Illustration 5. The average weekly for a group of 25 persons working in a factory was calculated to be Rs. 378.40. It was later discovered that one figure was misread as 160 instead of the correct value Rs. 200. Calculate average wage.

Solution. $\Sigma X = N\bar{X} = 25 \times 378.4 = 9460$

Less: Incorrect figure	160
	9300
Add: Correct figure	200
Total	9500

$\therefore$ *Correct* ΣX = 9500

Hence correct average $= \frac{9500}{25} = 380.$

(*b*) The mean of 200 observations was 50. Later on, it was discovered that two observations were wrongly read as 92 and 8 instead of 192 and 88. Find out the correct mean.

Solution.

Here $\bar{X} = 50$ and $N = 200$

$\Sigma X = 200 \times 50 = 10{,}000$

Less Incorrect observations	100
	9,900

Add Correct observation correct total = $\frac{280}{10,180}$

Correct mean = $\frac{10,180}{200}$ = 50.9

Mathematical Properties of Arithmetic Mean

The important mathematical properties of arithmetic mean are :

1. The algebraic sum of the deviations of all the observations from arithmetic mean is always zero, *i.e.*, $\Sigma (X - \bar{X}) = 0$. This shall be clear from the following example :

X	$(X - \bar{X})$
10	− 20
20	− 10
30	0
40	+ 10
50	+ 20
$\Sigma X = 150$	$\Sigma (X - \bar{X}) = 0$

Here $\bar{X} = \frac{\Sigma X}{N} = \frac{150}{5} = 30$. When the sum of the deviations from the actual mean, *i.e.*, 30, is taken it comes out to be zero. It is because of this property that the mean is characterised as a *point of balance, i.e.*, the sum of the positive deviations from mean is equal to the sum of the negative deviations from mean.

2. The sum of the squared deviations of all the observations from arithmetic mean is minimum, that is, less than the squared deviations of all the observations from any other value than the mean. The following example would clarify the point :

X	$(X - \bar{X})$	$(X - \bar{X})^2$
2	− 2	4
3	− 1	1
4	0	0
5	+ 1	1
6	+ 2	4
$\Sigma X = 20$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 10$

The sum of the squared deviations is equal to 10 in the above case. If the deviations are taken from any other value, the sum of the squared deviations would be greater than 10. This is known as the least square property of the arithmetic mean and becomes the basis for defining the concept of standard deviation.

3. If we have the arithmetic mean and number of observations of two or more than two related groups, we can compute combined average of these groups by applying the following formula :

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$\bar{X}_{12}$ = Combined mean of the two groups.

$\bar{X}_1$ = Arithmetic mean of the first group.

$\bar{X}_2$ = Arithmetic mean of the second group.

N_1 = Number of observations in the first group.

N_2 = Number of observations in the second group.

The following example will illustrate the application of the above formula :

Illustration 6(*a*). There are two branches of a company employing 100 and 80 employees respectively. If arithmetic means of the monthly salaries paid by two branches are Rs. 4570 and Rs. 6750 respectively, find the arithmetic mean of the salaries of the employees of the company as a whole.

Solution. We should compute the combined mean. The formula is

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Given $N_1 = 100,\ \bar{X}_1 = 4570,\ N_2 = 80,\ \bar{X}_2 = 6750$

$$\therefore \quad \bar{X}_{12} = \frac{(100\times 4570)+(80\times 6750)}{100+80} = \frac{997000}{180} = 5538.89$$

If we have to find out the combined mean of three related groups, the above formula can be extended as follows :

$$\bar{X}_{123} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3}$$

Illustration 6. (*b*) The mean of marks in Statistics of 100 students of a class was 72. The mean of marks of boys was 75, while their number was 70. Find out the mean marks of girls in the class. *(MBA, Osmania Univ, 2006)*

Solution. We are given $N = 100,\ \bar{X}_{12} = 72,\ \bar{X}_1$, *i.e.*, mean marks of boys = 75, N_1 = number of boys = 70. We have to find out the mean marks of girls, *i.e.*, $\bar{X}_2$.

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$72 = \frac{70(75)+30\bar{X}_2}{70+30}$$

$$7200 = 5250+30\,\bar{X}_2$$

$$\bar{X}_2 = \frac{1950}{30} = 65$$

Hence mean marks of girls in the class = 65.

Illustration 6. (*c*) The mean age of a combined group of men and women is 30 years. If mean age of the group of men is 32 and that of the group of women is 25, find out the percentage of men and women in the group.

Solution. Let N_1 represent percentage of men and N_2 percentage of women so that $N_1+N_2 = 100$.

We are given $\bar{X}_{12} = 30$

$\bar{X}_1 = 32$ (mean age of group of men)

$\bar{X}_2 = 25$ (mean age of group of women)

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$30 = \frac{N_1(32) + N_2(25)}{100}$$

$$3000 = 32N_1+(100-N_1)25$$

$$32N_1 + 2500 - 25N_1 = 3000 \quad \text{or} \quad N_1 = 71.43$$

$$N_2 = 100 - 71.43 = 28.57$$

Hence the percentage of men and women is respectively 71.43 and 28.57.

Merits and Limitations of Arithmetic Mean

The arithmetic mean is the most popular average in practice. It is due to the fact that it possesses first six out of seven characteristics of a good average and no other average possesses such a large number of characteristics.

However, arithmetic mean is unduly affected by the presence of extreme values. Also in open-end frequency distribution, it is difficult to compute mean without making assumption regarding the size of the class-interval of the open-end classes. The arithmetic mean is usually neither the most commonly occurring value nor the middle value in a distribution and in extremely asymmetrical distribution, it is not a good measure of central tendency.

Weighted Arithmetic Mean

One of the limitations of the arithmetic mean discussed above is that it gives equal importance to all the observations. But there are cases where the relative importance of the different observations is not the same. When this is so, we compute weighted arithmetic mean. The terms 'weight' stands for the relative importance of the different observations. The formula for computing weighted arithmetic mean is :

$$\bar{X}_w = \frac{\Sigma WX}{\Sigma W}$$

where $\bar{X}_w$ represents the weighted arithmetic mean

X = The variable.

W = Weights attached to the variable X.

An important problem that arises while using weighted mean is regarding selection of weights. Weights may be either actual or arbitrary, *i.e.,* estimated. Needless to say, if actual-weights are available, nothing like this. However, in the absence of actual-weights, arbitrary or imaginary weights may be used. The use of arbitrary weights may lead to some error, but this is better than no weights at all. In practice, it is found that if weights are intelligently assigned keeping the phenomena in view, the error involved will be so small that it can be easily overlooked.

Weighted mean is specially useful in problems relating to the construction of index numbers and standardised birth and death rates.

Illustration 7. A contractor employs three types of workers—male, female and children. To a male worker he pays Rs. 200 per day, to a female worker Rs. 150 per day and to a child worker Rs. 100 per day. What is the average wage per day paid by the contractor?

Solution. The average wage is not the simple arithmetic mean, *i.e.,* $\frac{200+150+100}{3}$ = Rs. 150 per day. If we assume that the number of male, female and child workers is the same, this answer would be correct. For example, if we take 10 workers in each case then the average wage would be

$$\bar{X} = \frac{(10\times 200)+(10\times 150)+(10\times 100)}{10+10+10} = \frac{2000+1500+1000}{30} = \frac{4500}{30} = \text{Rs. } 150$$

However, the number of male, female and child workers employed is generally different. If we know how many workers of each type are employed by the contractor in question, nothing like this. However, in the absence of this we take assumed weights. Let us assume that the number of male, female and child workers employed is 20, 15 and 5, respectively. The average wage would be the weighted mean calculated as follows :

Wage per day (Rs.) X	*No. of workers* W	WX
200	20	4000
150	15	2250
100	5	500
	$\Sigma W = 40$	$\Sigma WX = 6750$

$$\bar{X}_w = \frac{\Sigma WX}{\Sigma W} = \frac{6750}{40} = 168.75$$

Hence the average wage per day paid by the contractor is Rs. 168.75 to all types of workers.

B. MEDIAN

The median is the measure of central tendency which appears in the "middle" of an ordered sequence of values. That is, half of the observations in a set of data are lower than it and half of the observations are greater than it.

As distinct from the arithmetic mean which is calculated from the *value of every* observation in the series, the median is what is called a *positional* average. The term 'position' refers to the

place of a value in a series. The place of the median in a series is such that an equal number of observations lie on either side of it. For example, if the income of five persons is Rs. 7000, 7200, 7500, 7600, 7800, then the median income would be Rs. 7500. Changing any or both of the first two values with any other numbers with value of 7500 or less and/or changing any of the last two values to any other values with values of 7500 and more, would not affect the value of the median which would remain 7500. In contrast, in case of arithmetic mean the change in value of single observation would cause the value of the mean to be changed. Median is thus the central value of the distribution or the value that divides the distribution into two equal parts. If there are even number of observations in a series, there is no actual value exactly in the middle of the series and as such the median is indeterminate. In this case, the median is arbitrarily taken to be halfway between the two middle observations. For example, if there are 10 observations in a series, the median position is 5.5, that is the median value is halfway between the value of the observations that are 5th and 6th in order of magnitude. Thus when N is odd, the median is in an actual value with the remainder of the series in two equal parts on either side of it. If N is even, then the median is a derived figure, *i.e.*, half the sum of two values.

Calculation of Median—Ungrouped Data

Arrange the data in ascending or descending order of magnitude. (Both arrangements would give the same answer.)

Apply the formula : Median = Size of $\frac{N+1}{2}$th observation.

Illustration 8. From the following data of wages of 7 workers, compute the median wage :

Wages (in Rs.) 4600 4650 4580 4690 4660 4606 4640

Solution :

CALCULATION OF MEDIAN

S. No.	*Wages arranged in ascending order*
1	4580
2	4600
3	4606
4	4640
5	4650
6	4660
7	4690

Median = Size of $\frac{N+1}{2}$th observation = $\frac{7+1}{2}$ = 4th observation.

Value of 4th observation is 4640. Hence median wages = Rs. 4640.

In the above illustration, the number of observations was odd and, therefore, it was possible to determine the value of 4th observation. When the number of observations is even, for example, if in the above case the number of observations are 8 the median would be the value of $\frac{8+1}{2}$ = 4.5th observation. For finding out the value of 4.5th observation, we shall take the average of 4th and 5th observations. Hence the median shall be

$$\frac{4640+4650}{2} = 4645$$

Calculation of Median—Grouped Data

Determine the particular class in which the value of median lies. Use $\frac{N}{2}$ to locate the median class and not $\frac{N+1}{2}$ because in the use of grouped data it is $N/2$ which divides the area of the curve into two equal parts.

Apply the following formula for determining the exact value of median :

$$\text{Median} = L + \frac{N/2 - p.c.f.}{f} \times i$$

L = Lower limit of median class, *i.e.*, the class in which the middle observation in the distribution lies.

$p.c.f.$ = Preceding cumulative frequency to the median class.

f = Frequency of the median class.

i = The class-interval of the median class.

Illustration 9. (*a*) 1,500 workers are working in an industrial establishment. Their age is classified as follows :

Age (yrs.)	*No. of workers*	*Age (yrs.)*	*No. of workers*
18–22	120	38–42	184
22–26	125	42–46	162
26–30	280	46–50	86
30–34	260	50–54	75
34–38	155	54–58	53

Calculate the median age.

Solution :

CALCULATION OF MEDIAN AGE

Age group	*f*	*c.f.*
18–22	120	120
22–26	125	245
26–30	280	525
30–34	260	785
34–38	155	940
38–42	184	1,124
42–46	162	1,286
46–50	86	1,372
50–54	75	1,447
54–58	53	1,500

Median = Size of $\frac{N}{2}$ th observation = $\frac{1,500}{2}$ = 750th observation.

Hence median lies in the class 30–34.

$$\text{Median} = L + \frac{N/2 - p.c.f}{f} \times i = 30 + \frac{750 - 525}{260} \times 4 = 30 + 3.46 = 33.46$$

Hence the median age of the workers is 33.46 years.

(*b*) Calculate the median from the following data pertaining to the profits (in crore Rs.) of 125 companies :

Profits (Rs. crore)	*No. of companies*
less than 10	4
less than 20	16
less than 30	40
less than 40	76
less than 50	96
less than 60	112
less than 70	120
less than 80	125

(MBA, MD Univ., 2000)

Solution :

CALCULATION OF MEDIAN

Profits (Rs. Crore)	*No. of companies (f)*	*c.f.*
0 — 10	4	4
10 — 20	12	16
20 — 30	24	40
30 — 40	36	76
40 — 50	20	96
50 — 60	16	112
60 — 70	8	120
70 — 80	5	125

Med. = Size of $\frac{N}{2}$ th observation = $\frac{125}{2}$ = 62.5th observation.

Median lies in the class 30—40.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i$$

$$L = 30, N/2 = 62.5, p.c.f. = 40, f = 36, i = 10$$

$$\therefore \quad \text{Med.} = 30 + \frac{62.5 - 40}{36} \times 10 = 30 + 6.25 = 36.25.$$

Hence 50% of the companies have profits upto Rs. 36.25 crores and the remaining 50% of the companies have profits more than Rs. 36.25 crores.

Merits and Limitations of Median

The median is superior to arithmetic mean in certain respects. For example, it is especially useful in case of open-end distribution and also it is not influenced by the presence of extreme values. In fact when extreme values are present in the data, the median is a more satisfactory measure of central tendency than the mean.

The sum of the deviations of observations from median (ignoring signs) is *minimum*. In other words, the absolute deviation of observations from the median is less than from any other value in the distribution. For example, the median of items 4, 6, 8, 10 and 12 is 8. The deviations from 8 ignoring signs are 4, 2, 0, 2, 4 and the total is 12. This total will be smaller than the one obtained if deviations are taken from any other value. Thus, if deviations are taken from 7, the deviations ignoring signs would be 3, 1, 1, 3, 5 and the total is 13. In an estimation situation, if one is interested in minimising the absolute amount of error and the sign of the error is not particularly important, then the median is preferable to arithmetic mean.

However, since median is a positional average, its value is not determined by each and every observation. Also median is not capable of algebraic treatment. For example, median cannot be used for determining the combined median of two or more groups. Also the median is less reliable average than the mean for estimation purposes since it is more affected by sampling fluctuations. Furthermore, the median tends to be rather unstable value if the number of observations is small.

Related Positional Measures or Quantities

Besides median, there are other measures which divide a series into a equal number of parts. Important amongst these are quartiles, deciles and percentiles. These quartiles, deciles and percentiles are all special cases of *quantities*. Quartiles are those values of the variate which divide the total frequency into four equal parts, deciles divide the total frequency in 10 equal parts and the percentiles divide the total frequency in 100 equal parts. Just as one point divides a series into two parts, three points would divide it into four parts, 9 points into 10 parts and 99 points into 100 parts consequently there are only 3 quartiles, 9 deciles and 99 percentiles for a series. The quartiles are denoted by symbol Q, deciles by D and percentiles by P. The subscripts 1, 2, 3, etc., beneath Q, D and P would refer to the particular value that we want to compute. Thus Q_1 would denote first quartile Q_2 second quartile, Q_3 third quartile, D_1 first decile, D_8 eighth decile, P_1 first percentile, etc.

Graphically any set of these partition values serves to divide the area of the frequency curve or histogram into equal parts. If vertical lines are drawn at each quartile, for example, the area of the histogram will be divided by these lines into four equal parts. The 9 deciles divide the area of the histogram or frequency curve into 10 equal parts and the 99 percentiles divide the area into 100 equal parts.

In economics and business, quartiles are more widely used than deciles and percentiles. The quartiles are the points on the X-scale that divide the distribution into four equal parts. Obviously there are three quartiles, the second coinciding with the median. More precisely stated, the lower quartile, Q_1 is that point on the X-scale such that one-fourth of the total frequency is less than Q_1 and three-fourths is greater than Q_1. The upper quartile, Q_3, is that point on the X-scale such that three-fourths of the total frequency is below Q_3 and one-fourth is above it.

The deciles and percentiles are important in psychological and educational statistics concerning grades, rates, scores and ranks ; they are of use in economics and business in personnel department, productivity ratings and other situations.

Computation of Quartiles, Deciles, Percentiles, etc.

The procedure for computing quartiles, deciles, etc., is the same as for median.

For grouped data, the following formulae are used for quartiles, deciles and percentiles ;

$$Q_j = L + \frac{\frac{jN}{4} - p.c.f.}{f} \times i \qquad \text{for } j = 1, 2, 3$$

$$D_k = L + \frac{\frac{kN}{10} - p.c.f.}{f} \times i \qquad \text{for } k = 1, 2, \ldots, 9$$

$$P_l = L + \frac{\frac{lN}{100} - p.c.f.}{f} \times i \qquad \text{for } l = 1, 2, \ldots, 99$$

where the symbols have their usual meanings and interpretation.

Illustration 10. The profits earned by 100 companies during 2009-10 are given below :

Profits (Rs. lakhs)	*No. of companies*	*Profits (Rs. lakhs)*	*No. of companies*
20 — 30	4	60 — 70	15
30 — 40	8	70 — 80	10
40 — 50	18	80 — 90	8
50 — 60	30	90 — 100	7

Calculate Q_1, median, D_4 and P_{80} and interpret the values. *(MBA, D.U., 2001)*

Solution.

CALCULATION OF Q_1, Q_2, D_4 AND P_{80}

Profits (Rs. lakhs)	f	*c.f.*
20 – 30	4	4
30 – 40	8	12
40 – 50	18	30
50 – 60	30	60
60 – 70	15	75
70 – 80	10	85
80 – 90	8	93
90 – 100	7	100

Q_1 = Size of $N/4$th observation = $\frac{100}{4}$ = 25th observation.

Hence Q_1 lies in the class 40 — 50.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 40 + \frac{25-12}{18} \times 10 = 40 + 7.22 = 47.22$$

25 per cent of the companies earn an annual profit of Rs. 47.22 lakhs or less.

Median or Q_2 = Size of $\frac{2N}{4}$th observation = $\frac{200}{4}$ = 50th observation. Q_2 lies in the class 50—60.

$$Q_2 = L + \frac{2N/4 - p.c.f.}{f} \times i = 50 + \frac{50-30}{30} \times 10 = 50 + 6.67 = 56.67$$

50 per cent of the companies earn an annual profit of Rs. 56.67 lakhs or less.

D_4 = Size of $\frac{4N}{10}$th observation = 40th observation

D_4 lies in the class 50—60.

$$D_4 = L + \frac{4N/10 - p.c.f.}{f} \times i = 50 + \frac{40-30}{30} \times 10 = 50 + 3.33 = 53.33.$$

Thus 40 per cent of the companies earn an annual profit of Rs. 53.33 lakhs or less.

P_{80} = Size of $\frac{80N}{100}$th observation = $\frac{80 \times 100}{100}$ = 80th observation

P_{80} lies in the class 70—80.

$$P_{80} = L + \frac{80N/100 - p.c.f.}{f} \times i = 70 + \frac{80-75}{10} \times 10 = 70 + 5 = 75$$

This means that 80 per cent of the companies earn an annual profit of Rs. 75 lakhs or less and 20 per cent of the companies earn an annual profit of more than Rs. 75 lakhs.

Determination of Median, Quartiles, etc., Graphically

Median can be determined graphically by applying any of the following two methods :

1. Draw two ogives — one by 'less than' method and other by 'more than' method. From the point where both these curves intersect each other, draw a perpendicular on the X-axis. The point where this perpendicular touches the x-axis gives us the value of median.

2. Draw only one ogive by 'less than' method. Take the variable on the X-axis and frequency on the Y-axis. Determine the median value by the formula : median = Size of $\frac{N}{2}$th item. Locate this value on the Y-axis and from it draw a perpendicular on the cumulative frequency curve. From the point where it meets the ogive draw another perpendicular on the X-axis and the point where it meets the X-axis is the median.

The other partition values like quartiles, deciles and percentiles can also be determined graphically.

Illustration 11. Using the data of illustration 10, determine graphically the values of Q_1, Q_2, D_{40} and P_{80}.

Solution. Draw the ogive by the 'less than' method as shown in the graph.

To determine different quartiles, horizontal lines (broken) are drawn from the cumulative frequency values. For example, if we want to determine the value of median, a horizontal line can be drawn from the cumulative frequency value of 0.50 to the less than curve and then extending a vertical line to the horizontal axis. In a similar manner other values can be determined as shown in the graph. Therefore, $Q_1 = 47.22$, $Q_2 = 56.67$, $D_{40} = 53.33$ and $P_{80} = 75$. This may be noted down here that these graphical values are same as obtained by the formulae.

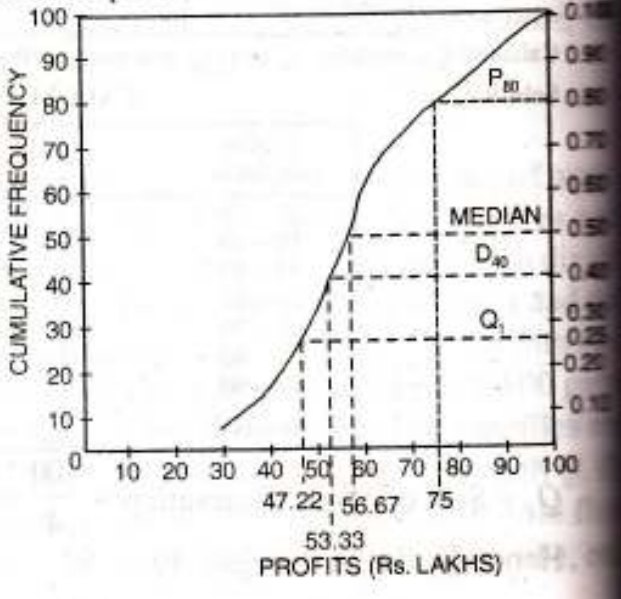


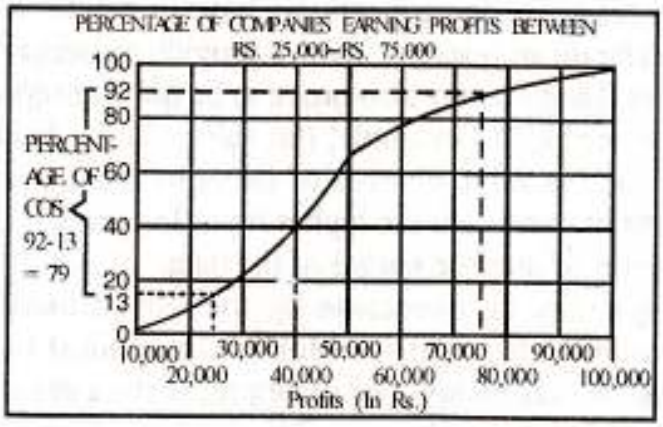
Illustration 12. You are given the net profits earned by some companies. Estimate graphically the percentage of companies getting profits between Rs. 25,000 and Rs. 75,000.

Profits (in Rs.)	*No. of companies*	*Profits (in Rs.)*	*No. of companies*
10,000—20,000	15	60,000—70,000	22
20,000—30,000	35	70,000—80,000	12
30,000—40,000	47	80,000—90,000	11
40,000—50,000	68	90,000—1,00,000	8
50,000—60,000	32		

Solution. Finding percentage from the given data :

Profits less than	*No. of companies*	*Percentage*
Rs. 20,000	15	6.0
" 30,000	50	20.0
" 40,000	97	38.8
" 50,000	165	66.0
" 60,000	197	78.8
" 70,000	219	87.6
" 80,000	231	92.4
" 90,000	242	96.8
" 1,00,000	250	100.0

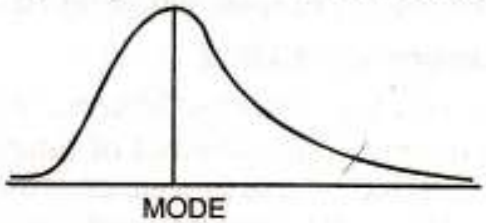
Plotting the data on the graph paper :



The graph shows clearly that the percentage of companies earning profits less than Rs. 75,000 is 92 and the percentage of companies earning profits less than Rs. 25,000 is 13. Thus the percentage of companies making profits between Rs. 25,000 and Rs. 75,000 is (92–13) = 79.

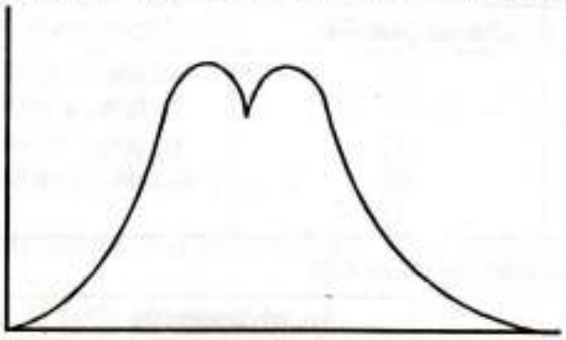
C. MODE

Mode is defined as that value which occurs the maximum number of times, *i.e.*, having the maximum frequency. For example, if we take the values of six different observations as 5, 8, 10, 8, 5, 8, mode will be 8 as it has occurred maximum number of times, *i.e.*, 3 times. Graphically, it is the value on the *X*-axis below the peak, or highest point, of the frequency curve as can be seen from the following diagram.



This interpretation of the statistical mode is analogous to that of the fashion mode. A person dressing with current styles is "in the mode". But a current fashion can be a poor description of what most persons are wearing because of the variety of styles worn by the general public. In statistics, the mode only tells us which single value occurs most often; it may, therefore, represent a majority of the total population.

It is possible that a distribution may be bimodal. This happens when there may be two or more values of equal or nearly equal occurrence as can be seen from the following diagram :



The presence of more than one mode has a special significance in statistical analysis, for it indicates potential trouble. It is usually dangerous to compare bimodal populations or to draw conclusions about them because they usually arise when there is some non-homogeneous factor present in the population.

If the collected data produce a bimodal distribution, the data themselves should be questioned. Quite often such a condition is caused by the taking of too small a sample; the difficulty can be remedied by increasing the sample size. In instances where a distribution is bimodal and nothing can be done to change it, the mode is obviously eliminated as a measure of central tendency.

There are many situations in which arithmetic mean and median fail to reveal the true characteristic of data. For example, when we talk of most common wage, most common income, most common height, most common size of shoe or ready-made garments, we have in mind mode and not the arithmetic mean or median discussed earlier. The mean does not always provide an accurate reflection of the data due to the presence of extreme values. Median may also prove to be quite unrepresentative of the data owing to an uneven distribution of the series. For example, the values in the lower half of a distribution range from, say, Rs. 10 to 100, while the same number of items in the upper half of the series range from Rs. 100 to Rs. 6,000 with most of them near the higher limit. In such a distribution the median value Rs. 100 will provide little indication of the true nature of the data.

Both these shortcomings may be overcome by the use of mode. Mode refers to that value which occurs most frequently in a distribution. Mode is the easiest to compute since it is the value corresponding to the maximum frequency. For example, if the data is :

Size of shoes	:	5	6	7	8	9	10	11
No. of persons	:	10	20	25	40	22	15	6

the modal size is '8' since it appears maximum number of times in the data.

Calculation of Mode

Determining the precise value of the mode of a frequency distribution is by no means an elementary calculation. Essentially it involves fitting mathematically some appropriate type of frequency curve to the grouped data and the determination of the value on the X-axis below the peak of the curve. However, there are several elementary methods of *estimating* the mode. These methods have been discussed for ungrouped and grouped data.

Calculation of Mode—Ungrouped Data

For determining mode count, the number of observations the various values repeat themselves, and the value which occurs the maximum number of times is the modal value.

Illustration 13. The following figures relate to the preferences with regard to size of screen (in inches) of T.V. sets of 30 persons selected at random from a locality. Find the modal size of the T.V. screen.

12	20	12	24	29
20	12	20	29	24
24	20	12	20	24
29	24	24	20	24
24	20	24	24	12
24	20	29	24	24

Solution. CALCULATION OF MODAL SIZE

Size in inches	*Tally*	*Frequency*
12	𝍸	5
20	𝍸 |||	8
24	𝍸 𝍸 |||	13
29	||||	4
	Total	30

Since size 24 occurs the maximum number of times, therefore, the modal size of T.V. screen is 24 inches.

Calculation of Mode—Grouped Data

In the case of grouped data, the following formula is used for calculating mode :

$$Mo = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \qquad ...(i)$$

where L = Lower limit of the modal class.

Δ_1 = The difference between the frequency of the modal class and the frequency of the pre-modal class, *i.e.*, preceding class.

Δ_2 = The difference between the frequency of the modal class and the frequency of the post-modal class, *i.e.*, succeeding class.

i = The size of the modal class.

Another form of this formula is :

$$Mo = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \qquad ...(ii)$$

where L = Lower limit of the modal class.

f_1 = Frequency of the modal class.

f_0 = Frequency of the class preceding the modal class.

f_2 = Frequency of the class succeeding the modal class.

While applying the above formula for calculating mode, it is necessary to see that the class intervals are *uniform* throughout. If they are unequal, they should first be made equal on the assumption that the frequencies are equally distributed throughout the class, otherwise we will get misleading results.

A distribution having only one mode is called *unimodal.* If it contains more than one mode, it is called *bimodal* or *multimodal.* In the latter case, the value of mode cannot be determined by the above formula and hence mode is *ill-defined* when there is more than one value of mode.

Where mode is ill-defined, its value may be ascertained by the following approximate formula* based upon the relationship between mean, median and mode.

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

Illustration 14. The following data relate to the sales of 100 companies :

Sales (Rs. lakhs)	*No. of companies*	*Sales (Rs. lakhs)*	*No. of companies*
Below 60	12	66 — 68	10
60—62	18	68—70	3
62—64	25	70—72	2
64—66	30		

Calculate the value of modal sales.

*See page 99.

Solution. Since the maximum frequency 30 is in the class 64—66, therefore, 64—66 is the modal class.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 64,\ \Delta_1 = (30 - 25) = 5,\ \Delta_2 = (30 - 10) = 20,\ i = 2$$

$$\text{Mode} = 64 + \frac{5}{5 + 20} \times 2 = 64 + \frac{10}{25} = 64.4$$

Hence modal sales are Rs. 64.4 lakhs.

Locating Mode Graphically

In a frequency distribution the value of mode can also be determined graphically. The steps in calculation are :

1. Draw a histogram of the given data.
2. Draw two lines diagonally on the inside of the modal class bar, starting from each upper corner of the bar to the upper corner of the adjacent bar.
3. Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis (horizontal scale) which gives us modal value.

Illustration 15. The daily profits in rupees of 100 shops are given as follows :

Profits (in Rs. lakhs)	No. of shops	Profits (in Rs. lakhs)	No. of shops
0—100	12	300—400	20
100—200	18	400—500	17
200—300	27	500—600	6

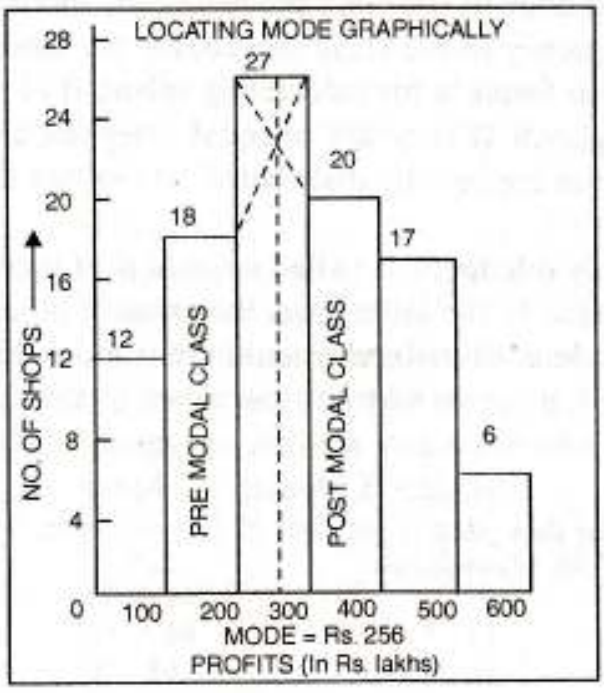
Draw the histogram and thence find the modal value. Check this value by direct calculation.

Solution.

Direct calculation :

Mode lies in the class 200—300.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 200 + \frac{9}{9 + 7} \times 100 = 256.25$$



From the above diagram, the modal value is also 256. Hence by both the methods we get the same value of mode.

Mode can also be determined from frequency polygon in which case perpendicular is drawn on the base from the apex of the polygon and the point where it meets the base gives the modal value.

However, graphic method of determining mode can be used only where there is one class containing the highest frequency. If two or more classes have the same highest frequency, mode cannot be determined graphically.

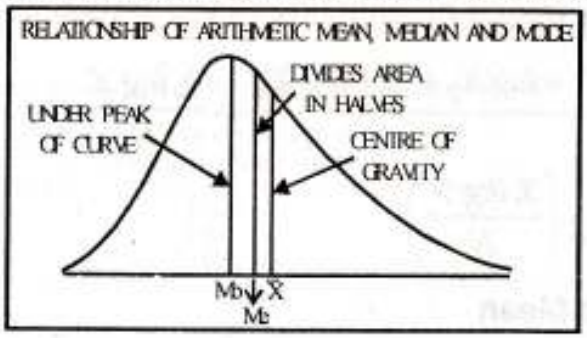
Merits and Limitations of Mode

Like median, the mode is not affected by extreme values and its value can be obtained in open-end distributions without ascertaining the class limits. Mode can be easily used to describe qualitative phenomenon. For example, when we want to compare the consumer preferences for different types of products, say, soap, toothpastes, etc., or different media of advertising, we should compare the modal preferences. In such distributions where there is an outstanding large frequency, mode happens to be meaningful as an average.

However, mode is not a rigidly defined measure as there are several formulae for calculating the mode, all of which usually give somewhat different answers. Also the value of mode cannot always be computed, such as, in case of bimodal distributions.

Relationship among Mean, Median and Mode

A distribution in which the values of mean, median and mode coincide is known as *symmetrical* distribution. Conversely stated, when the values of mean, median and mode are not equal, the distribution is known as *asymmetrical* or *skewed*. In moderately skewed or asymmetrical distributions, a very important relationship exists among mean, median and mode. In such distributions, the distance between the mean and the median is approximately one-thrid of the distance between the mean and mode as will be clear from the following diagram :



Karl Pearson has expressed this approximate relationship as follows :

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

or $$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean}$$

and $$\text{Median} = \frac{2\ \text{Mean} + \text{Mode}}{3}$$

If we know any of the two values out of the three, we can compute the third from these relationships. The following example will illustrate this point :

Illustration 16. In a moderately asymmetrical distribution the Mode and Mean are 32.1 and 35.4 respectively. Calculate the Median.

Solution. Mode = 3 Median –2 Mean

Mode = 32.1, Mean = 35.4

Substituting the values

32.1 = 3 Median –2 (35.4) or 3 Med = 102.9 or Med. = 34.3.

2. GEOMETRIC MEAN

In business and economic problems, very often we are faced with questions pertaining to percentage rates of change over time. Neither the mean, the median nor mode is the appropriate average to use in these instances. For example, consider the following figures of sale of a company :

Year :	2007	2008	2009	2010
Sales (million tonnes) :	20.2	22.5	23.9	28.0

Suppose we want to find out the average percentage rate of change per year in sales. To answer this question we must specify what we mean by the 'Average percentage rate of change per year'. The most generally useful interpretation of this term is the constant percentage rate of change which if applied each year would take us from the first to the last figure. Hence in the above illustration we would be interested in that constant yearly percentage rate of change which would be required to move from 20.2 million tonnes of sales in 2007 to 28.0 million tonnes in 2010. None of the previously discussed averages provides the correct answer to this question. The correct answer can be obtained through the use of the geometric mean or, what amounts to the same thing, through the use of the familiar compound interest formula. In the discussion which follows, the geometric mean is defined, and the relationship between this average and compound interest calculations is indicated.

Geometric mean is defined as the Nth root of the product of N observations of a given data. If there are two observations, we take the square root; if there are three observations, the cube root; and so on, symbolically.

$$G.M. = \sqrt[N]{X_1 \times X_2 \times X_3 \times ..... \times X_N}$$

where $X_1, X_2, X_3 ..., X_N$, refer to the various observations of the data.

When the number of observations is three or more the task of multiplying the number and of extracting the root becomes quite difficult. To simplify calculations logarithms are used. Geometric mean is then calculated as follows :

$$\log \text{G.M.} = \frac{\log X_1 + \log X_2 + ... + \log X_N}{N} = \frac{\Sigma \log X}{N}$$

$$\therefore \quad \text{G.M.} = \text{antilog} \left(\frac{\Sigma \log N}{N} \right).$$

Calculation of Geometric Mean

In ungrouped data, geometric mean is calculated with the help of the following formula :

$$\text{G.M.} = \text{A.L.} \left(\frac{\Sigma \log X}{N} \right)$$

In grouped data, for calculating geometric mean first we will find the midpoints and then apply the following formula :

$$\text{G.M.} = \text{A.L.} \left(\frac{\Sigma f \log X}{N} \right)$$

where X = midpoint.

Compound Interest Formula

The compound interest formula is expressed as follows :

$$P_n = P_0 (1 + r)^n$$

where P_n = amount accumulated at the end of n periods,

P_0 = Original principal,

r = Rate of interest expressed as a decimal, and

n = Number of compound periods.

It follows from the above formula that :

$$r = \sqrt[n]{\frac{P_n}{P_0}} - 1$$

If interest is compounded at different rates in each time period, and if these successive rates are denoted by $r_1, r_2, \ldots, r_n$, then the amount accumulated at the end of n periods with an original principal of P_0 is

$$P_n = P_0 (1 + r_1)(1 + r_2) \ldots\ldots (1 + r_n).$$

Applications of Geometric Mean

Geometric mean is specially useful in the following cases :

1. The geometric mean is used to find the average per cent increase in sales, production, population or other economic or business data. For example, from 2008 to 2010 prices increased by 5%, 10% and 15% respectively. The average annual increase is not 11% as given by the arithmetic average but 10.9% as obtained by the geometric mean. This average is also useful in measuring the growth of population, because population increases in geometric progression.

2. Geometric mean is theoretically considered to be the best average in the construction of index number.* It makes index numbers satisfy the time reversal test and gives equal weights to equal ratio of change.

3. It is an average which is most suitable when large weights have to be given to small values of observations and small weights to large values of observations, situations which we usually come across in social and economic fields.

The following examples illustrate the use of geometric mean.

Illustration 17. Compared to the previous year the overhead expenses went up by 32% in 2008; they increased by 4% in the next year and by 50% in the following year. Calculate the average rate of increase in the overhead expenses over the three years.

Solution. In average ratios and percentages, geometric mean is more appropriate. Applying geometric mean here :

% Rise	*Expenses at the end of the year taking preceding year as 100*	*log X*
32	132	2.1206
40	140	2.1461
50	150	2.1761
		$\Sigma \log X = 6.4428$

$$\text{G.M.} = \text{A.L.}\left(\frac{\Sigma \log X}{N}\right) = \text{A.L.}\left(\frac{6.4428}{3}\right) = \text{A.L. } 2.1476 = 140.5.$$

Average rate of increase in overhead expenses

$= 140.5 - 100 = 40.5\%$.

Illustration 18. The annual rates of growth of output of a factory in 5 years are 5.0, 7.5, 2.5, 5.0 and 10.0 respectively. What is the compound rate of growth of output per annum for the period ?

Solution.

CALCULATING COMPOUND RATE OF GROWTH

Annual rate of growth	*Output relatives at the end of the year*	*log X*
5.0	105.0	2.0212
7.5	107.5	2.0314
2.5	102.5	2.0107
5.0	105.0	2.0212
10.0	110.0	2.0414
		$\Sigma \log X = 10.1259$

$$\text{G.M.} = \text{A.L.}\left(\frac{\Sigma \log X}{N}\right) = \text{A.L.}\left(\frac{10.1259}{5}\right) = \text{A.L. } 2.0252 = 105.9.$$

The compound rate of growth of output per annum for the period is $105.9 - 100 = 5.9\%$.

*Please refer to chapter on Index Numbers.

Illustration 19. A piece of property was purchased for Rs. 2,00,000 and sold 10 years later for Rs. 23,26,000. What is the average annual rate of return on the original investment ?

Solution. $2,00,000\ X^{10} = 3,26,000$

$$X^{10} = \frac{3,26,000}{2,00,000} = 1.63$$

$$\text{Log } X = \frac{\text{Log } 1.63}{10} = \frac{0.2122}{10} = 0.0212$$

X = A.L. (0.0212) = 1.05 or 105%.

Hence the investment yielded a mean rate return of 105 – 100 = 5 per cent over the 10-year period.

Combined Geometric Mean

Just as we have talked of combined arithmetic mean, in a similar manner we can also talk of combined geometric mean. If the geometric mean of N observations is 6 and these N observations are divided into two sets first containing N_1 and second containing N_2 observations having G_1 and G_2 as the respective geometric means, then

$$\text{Log } G = \frac{N_1 \text{ Log } G_1 + N_2 \text{ Log } G_2}{N_1 + N_2}$$

Thus if the geometric mean of 5 observations is 20 and of another 10 observations is 35.28, the combined geometric mean shall be

$$\text{Log } G = \frac{5 \text{ Log } 20 + 10 \text{ Log } 35.28}{5 + 10} = \frac{(5 \times 1.3010) + (10 \times 1.5475)}{15}$$

$$= \frac{6.505 + 15.475}{15} = \frac{21.98}{15} = 1.465$$

$$\therefore \qquad G = \text{A.L. } 1.465 = 29.17.$$

Illustration 20. Three groups of observations contain 8, 7, and 5 observations. Their geometric means are 8.52, 10.12 and 7.75 respectively. Find the geometric mean of the 20 observations in the single group formed by pooling the three groups.

Solution.

$$\text{Log } G = \frac{N_1 \text{ Log } G_1 + N_2 \text{ Log } G_2 + N_3 \text{ Log } G_3}{N_1 + N_2 + N_3}$$

$$= \frac{8 \text{ Log } 8.52 + 7 \text{ Log } 10.12 + 5 \text{ Log } 7.75}{8 + 7 + 5}$$

$$= \frac{(8 \times .9304) + (7 \times 1.0052) + (5 \times .8893)}{20}$$

$$= \frac{7.4432 + 7.0364 + 4.4465}{20} = \frac{18.9261}{20} = 0.9463$$

$$G = \text{A.L. } 0.9463 = 8.837.$$

Hence the combined geometric mean of the 20 observations taken together is 8.837.

Merits and Limitations of Geometric Mean

Geometric mean is highly useful in averaging ratios and percentages and in determining rates of increase and decrease. It is also capable of algebraic manipulation. For example, if the geometric mean of two or more series and their numbers of observations are known, a combined geometric mean can easily be calculated.

However, compared to arithmetic mean, this average is more difficult to compute and interpret. Also geometric mean cannot be computed when there are both negative and positive values in a series or more observations are having zero value.

E. HARMONIC MEAN

The harmonic mean is based on the reciprocal of the numbers averaged. It is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observation. Thus by definition

$$\text{H.M.} = \frac{N}{\left(\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}+...+\frac{1}{X_N}\right)}$$

When the number of observations is large, the computation of harmonic mean in the above manner becomes tedious. To simplify calculations, we obtain reciprocals of the various observations and apply the following formulae :

For ungrouped data, $\text{H.M.} = \frac{N}{\Sigma\left(\frac{1}{X}\right)}$

For grouped data, $\text{H.M.} = \frac{N}{\Sigma\left(f \times \frac{1}{X}\right)}$ or $\frac{N}{\Sigma\left(\frac{f}{X}\right)}$*

Illustration 21. (*i*) Calculate harmonic mean of numbers 10, 20 25, 40, 50. (*ii*) Calculate harmonic mean from the following frequency distribution :

X :	0—10	10—20	20—30	30—40	40—50
f :	8	15	20	4	3

Solution. (*i*)

CALCULATION OF HARMONIC MEAN

X	$1/X$
10	0.100
20	0.050
25	0.040
40	0.025
50	0.020
	$\Sigma\ 1/X = 0.235$

$$\text{H.M.} = \frac{N}{\Sigma\left(\frac{1}{X}\right)} = \frac{5}{0.235} = 21.28$$

(*ii*)

CALCULATION OF HARMONIC MEAN

Variable	X	f	$f \times 1/X$
0–10	5	8	1.600
10–20	15	15	1.000
20–30	25	20	0.800
30–40	35	4	0.114
40–50	45	3	0.067
		$N = 50$	$\Sigma\left(f \times \frac{1}{X}\right) = 3.581$

$$\text{H.M.} = \frac{N}{\Sigma\left(f \times \frac{1}{X}\right)} = \frac{50}{3.581} = 13.96.$$

*There is no need to first calculate $1/X$ and then multiply it by f. We can directly obtain f/X to simplify calculation.

Applications of Harmonic Mean

The harmonic mean is restricted in its field of applications.* It is useful for computing the average rate of increase of profits or average speed at which a journey has been performed or the average price at which an article has been sold. The rate usually indicates the relation between two different types of measuring units that can be expressed reciprocally. For example, if a man walked 20 km., in 5 hours, the rate of his walking speed can be expressed as follows:

$$\frac{20 \text{ km.}}{5 \text{ hours.}} = 4 \text{ km. per hour}$$

where the unit of the first term is a km., and the unit of the second term is an hour. Reciprocally,

$$\frac{5 \text{ hours}}{20 \text{ km.}} = \frac{1}{4} \text{ hour per km.}$$

where the unit of the *fitst term is an hour* and the unit of the *second term is a kilometre*.

Illustration 22. In a certain factory a unit of work is completed by A in 4 minutes, by B in 5 minutes, by C in 6 minutes, by D in 10 minutes and by E in 12 minutes.

(a) What is the average number of units of work completed per minute ?

(b) At this rate how many units will they complete in a six-hour day ?

Solution. (a) The average number of units per minute will be obtained by calculating the harmonic mean.

CALCULATION OF HARMONIC MEAN

X	$1/X$
4	0.250
5	0.200
6	0.167
10	0.100
12	0.083
	$\Sigma\ 1/X = 0.8$

$$\text{H.M.} = \frac{N}{\Sigma\left(\frac{1}{X}\right)} = \frac{5}{0.8} = 6.25$$

Hence the average number of units completed per minute = 6.25.

The average units per minute = $\frac{1}{6.25} = 0.16$.

In 6 hours, *i.e.*, 360 minutes, total number of units produced will be $360 \times 0.16 \times 5 = 288$ by all the five workers combined.

Illustration 23. A toy factory has assigned a group of 4 workers to complete an order of 1,400 toys of a certain type. The productive rates of the four workers are given below :

Workers	*Productive rates*
A	4 minutes per toy
B	6 minutes per toy
C	10 minutes per toy
D	15 minutes per toy

Find the average minutes per toy by the group of workers.

Solution. If we assume that each of the four workers is assigned the same number of toys (constant value) to meet the orders, or $\frac{1,400}{4} = 350$ toys per worker, the arithmetic mean would give the correct answer.

$$\bar{X} = \frac{4 + 6 + 10 + 15}{4} = \frac{35}{4} = 8\frac{3}{4} \text{ minutes per toy}$$

*The harmonic mean is a measure of central tendency for data expressed as rates, such as kms., per hour, tons per day, quantity per litre, miles per fortnight, etc.

Verification

Time required by A to complete 350 toys = 350 × 4 = 1,400 minutes
Time required by B to complete 350 toys = 350 × 6 = 2,100 minutes
Time required by C to complete 350 toys = 350 × 10 = 3,500 minutes
Time required by D to complete 350 toys = 350 × 15 = 5,250 minutes
12,250 minutes.

In 12,250 minutes 1,400 toys will be completed.
Hence, in completing one toy time taken will be

$$\frac{12,250}{1,400} = 8\frac{3}{4} \text{ minutes.}$$

However, if we assume that each worker works the same amount of time but produces different number of toys, harmonic mean would be more appropriate. This assumption is more true in practice (people working same amount of time but having different output).

$$\text{H.M.} \quad \frac{4}{\frac{1}{4}+\frac{1}{6}+\frac{1}{10}+\frac{1}{15}} = \frac{4 \times 60}{35} = 6\frac{6}{7} \text{ minutes per toy}$$

$$\text{Time required to complete 1,400 toys} = \frac{1,400 \times 48}{7} = 9,600 \text{ minutes}$$

Verification. Each worker works for $\frac{9,600}{4} = 2,400$ minutes

Toys produced by A in 2400 minutes $= \frac{2400}{4} = 600$

" " " B " " " $= \frac{2400}{6} = 400$

" " " C " " " $= \frac{2400}{10} = 240$

" " " D " " " $= \frac{2400}{15} = 160$

Total = 1,400

Merit and Limitations of Harmonic Mean

The harmonic mean, like the arithmetic mean and geometric mean, is computed from all observations. It is useful in special cases for averaging rates.

However, harmonic mean cannot be computed when there are both positive and negative observations or one or more observations have zero value. It also gives largest weight to smallest observations and as such is not a good representation of a statistical series. In dealing with business problems, harmonic mean is rarely used.

Relationship among the Averages

In any distribution when the original observations differ in size, the values of A.M., G.M. and H.M. would also differ and will be in the following order :

$$\text{A.M.} \geq \text{G.M.} \geq \text{H.M.}$$

i.e., arithmetic mean is greater than geometric mean and geometric mean is greater than harmonic mean. The equality signs hold only if all the observations $X_1, X_2, \ldots\ldots X_N$ are identical.

Progressive Average

This average is also based upon the arithmetic mean. The important features of this average are :

(*i*) It is a cumulative average. In the calculation of this average all previous figures are added and no previous figure is left as is done in the case of moving average.* The progressive average for the first year would remain the same ; the progressive average for the second year is equal to $\frac{a+b}{2}$; for the third year $\frac{a+b+c}{3}$, for the fourth year $\frac{a+b+c+d}{4}$; and so on.

(*ii*) The average value can be obtained for all the years. The moving average, on the other hand, cannot be computed for all the years. The longer the period of moving average, the greater the number of years for which the moving average cannot be computed.

This average is generally used during the early years of the working of a business. For example, the figures of sales, profits or production of each successive year may be compared with the respective figures for the entire previous period in order to find out how a business is growing.

The process of computing the progressive average shall be clear with the help of the following example :

Illustration 24. Calculate progressive average from the following data :

Year	*Sale of steel (in m. tonnes)*	*Year*	*Sale of steel (in m. tonnes)*
2004	12	2008	25
2005	14	2009	22
2006	15	2010	30
2007	18		

Solution.

CALCULATION OF PROGRESSIVE AVERAGE

Year	*Sales (in m. tonnes)*	*Progressive totals*	*Years included*	*Progressive average*
2004	12	12	1	12.00
2005	14	26	2	13.00
2006	15	41	3	13.67
2007	18	59	4	14.75
2008	25	84	5	16.80
2009	22	106	6	17.67
2010	30	136	7	19.43

The progressive average makes it clear that the above company is steadily progressing year after year.

Which Average to use ?

We have explained different methods of computing the various types of averages and also their distinctive features. At this point, the reader can question "which of these average should I use ?" Or "which of these is the best average to be used ?"

It must be clearly understood that no single average can be regarded as best for all purposes. The following two considerations should be kept in mind in the selection of an average :

1. The type of data available. Are they badly skewed (avoid the mean), gappy around the middle (avoid the median), or unequal in class-interval (avoid the mode) ?

2. The concept of the typical value required by the problem. Within the framework of descriptive statistics, the main requirement is to know what each average means and then select one that fulfils the purpose on hand. Is a composite average of all absolute or relative values

*For details please refer to chapter on Business Forecasting and Time Series Analysis.

geometric mean) ? Or, is a middle value needed (median), or the value (mode) ?

it may even be advisable to work out more than one average and present to be sure, this procedure creates an added burden for the reader as well as for the added burden is preferable to the use of a single average that may description. To use it alone is like looking through a keyhole : the part you can see cannot give a full idea of the whole room.

following cases, arithmetic mean should not be used :

skewed distributions.

distributions with open-end intervals.

the distribution is unevenly spread, concentration being small or large at irregular

an average rate of growth or change over a period of time is required.

the observations form a geometric progression, *i.e.*, 1, 2, 4, 8, 16, etc.

averaging rates (*i.e.*, speed, fluctuations in the prices of articles, etc.).

there are very large and very small values of observations arithmetic mean would misleading on account of undue influence of extreme values.

aside the above specific cases where either median, mode, geometric mean or mean is more appropriate, in other cases we should apply as a rule of thumb the mean—the most popular and widely used average in practice.

Median. The median is generally the best average in open-end grouped distributions, where if plotted as a frequency curve one gets a J or reverse *J* curve, for example, price distribution or income distribution. In such cases very high or very low values cause the mean to be higher or lower than the most "common" values. In such instances, median may be more representative to use in describing the mass of data.

Mode. Generally speaking, the significance of mode lies in the fact that it can be used to qualitative data. The mode can be used in problems involving the expression of where quantitative measurements are not possible. Thus the preferred type of design among a number of alternative designs would be the modal design. If we want consumer preferences for different kinds of products, or different kinds of advertising, we can compare the modal preferences expressed by different groups of people but we cannot the median or mean. Mode is a particularly useful average for discrete series, *e.g.*, number of people wearing a given size of shoes, or number of children per household, etc. The mode is best suited where there is an outstandingly large frequency.

Geometric Mean. The geometric mean is typically used in averaging index numbers, rates of change, ratios and other sets of data expressed in percentage form. It is particularly important in Economics and Business Statistics in index number construction.

Harmonic Mean. Harmonic mean is useful in problems in which values of a variable are compared with a constant quantity of another variable, *i.e.*, rates, time, distance covered within certain time and quantities purchased or sold per unit, etc.

GENERAL LIMITATIONS OF AN AVERAGE

1. Since an average is a single value representing a group of values, it must be properly interpreted, otherwise, there is every possibility of jumping to wrong conclusion. This can be best illustrated with the help of a story. A person had to cross a river from one bank to another. He was not aware of the depth

of the river, so he enquired from another man who told him that the average depth of water is 160 cms. The man was 175 cms and he thought that he can very easily cross the river because all the time he would be above the water level. So he started. In the beginning the level of water was very shallow but as he reached the middle, the water was 500 cms deep and he lost his life. The man was drowned because he had a misconception that average depth means uniform depth throughout. But it is not so. An average represents a group of values and lies somewhere in between the two extreme values.

2. An average may give us a value that does not exist in the data. For example, the arithmetic mean of 100, 300, 250, 50, 100 is $\frac{800}{5} = 160$, a value that does not exist in the data.

3. At times an average may give absurd results. For example, if we are calculating average size of a family we may get a value 4.8 . But this is impossible as persons cannot be in fractions. However, we should remember that it is an average value representing the entire group of families.

4. Measures of central value fail to give us any idea about the formation of the series. Two or more series may have the same central value but may differ widely in composition. For example, observe the following two series :

Series A :	150	170	190	210	280
Series B :	300	500	20	78	102

In both series, average $\bar{X} = 200$.

5. We must remember that an average is a measure of central tendency. Hence unless the data show a clear-cut concentration of observations an average may not be meaningful at all. This evidently precludes the use of any average to typify a bimodal or a U-shaped or a J-shaped distribution.

MISCELLANEOUS ILLUSTRATIONS

Illustration 25. Calculate mean, median and mode for the following data pertaining to marks in Statistics out of 140 marks for 80 students in a class :—

Marks more than :	0	20	40	60	80	100	120
No. of Students :	80	76	50	28	18	9	3

(MBA, K.U., 2002)

Solution. **CALCULATION OF MEAN, MEDIAN AND MODE**

Marks	*No. of Students* f	*m.p.* X	$(X-70)/20$ d	fd	*c.f.*
0—20	4	10	−3	−12	4
20—40	26	30	−2	−52	30
40—60	22	50	−1	−22	52
60—80	10	70	0	0	62
80—100	9	90	+1	+9	71
100—120	6	110	+2	+12	77
120—140	3	130	+3	+9	80
	$N = 80$			$\Sigma fd = -56$	

Mean : $\bar{X} = A + \frac{\Sigma fd}{N} \times i = 70 - \frac{56}{80} \times 20 = 70 - 14 = 56$

Median : Med = Size of $\frac{N}{2}$ th observation $= \frac{80}{2}$ = 40th observation Median lies in the class 40–60.

Median lies in the class 40—60

$$\text{Med} = L + \frac{N/2 - p.c.f}{f} \times i = 40 + \frac{40 - 30}{22} \times 20 = 40 + 9.09 = 49.09$$

Mode : Since the highest frequency is 26, mode lies in the class 20–40.

$$Mo = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 20 + \frac{22}{22 + 4} \times 20 = 20 + 16.92 = 36.92$$

Illustration 26. The following table gives the distribution of weekly wages of 600 workers of a factory :

Weekly wages (in Rs.)	*Frequency*	*Weekly wages (in Rs.)*	*Frequency*
Below 375	69	600—675	58
375—450	167	675—750	24
450—525	207	750—825	10
525—600	65		

(*a*) Draw an ogive for the above data and thence obtain the median value. Check it against calculated value.
(*b*) Obtain the limits of weekly wages of central 50 per cent of the workers. *(MBA, Delhi Univ., 1996)*

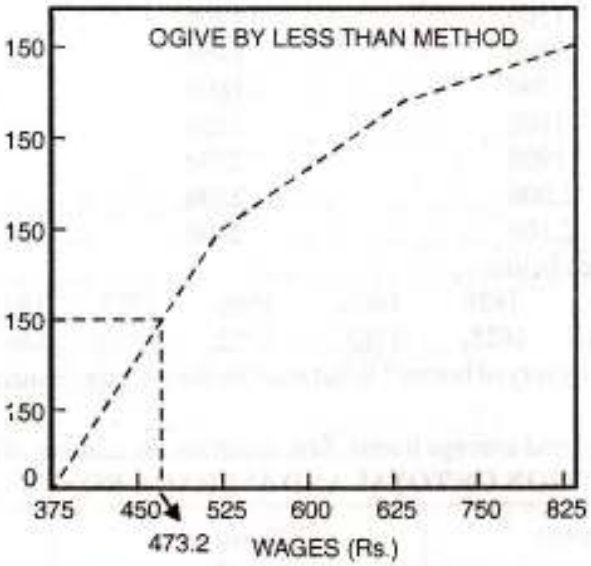
Solution. **CALCULATION OF MEDIAN**

Weekly wages (in Rs.)	*f*	*Cum freq.*
Less than 375	69	69
" " 450	167	236
" " 525	207	443
" " 600	65	508
" " 675	58	566
" " 750	24	590
" " 825	10	600

Med. = Size of $\frac{N}{2}$ th observation = $\frac{600}{2}$ = 300th observation

Median lies in the class 450 – 525.

$$\text{Med} = L + \frac{N/2 - p.c.f.}{f} \times i = 450 + \frac{300 - 236}{207} \times 75 = 450 + 23.2 = \text{Rs. } 473.2$$



The median value as shown in the graph above is also Rs. 473.2.

(*b*) The limits of weekly wages of central 50 per cent of the workers shall be given by Q_1 and Q_3.

Q_1 = Size of $\frac{N}{4}$ th observation = $\frac{600}{4}$ =150th observation

Q_1 lies in the class 375—450.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i$$

$$= 375 + \frac{150 - 69}{167} \times 75 = 375 + 36.38 = \text{Rs. } 411.38$$

Q_3 = Size of $\frac{3N}{4}$ th observation = $\frac{3 \times 600}{4}$ = 450th observation

Q_3 lies in the class 525–600.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i$$

$$= 525 + \frac{450 - 443}{65} \times 75 = 525 + 8.08 = \text{Rs. } 533.08.$$

Hence the limits of weekly wages of central 50 per cent workers are Rs. 411.38 and Rs. 533.08.

Illustration 27. In a factory, there are 100 skilled, 250 semi-skilled and 150 unskilled workers. It has been observed that on an average a unit length of a particular fabric is woven by a skilled worker in 3 hours, by semi-skilled workers in 4 hours and by an unskilled worker in 5 hours. After a training of 2 years, the semi-skilled workers are expected to become skilled and the unskilled workers to become semi-skilled. How much less time will be required after 2 years of training for weaving unit length of fabric by an average worker ?

Solution. Average time per worker before training is.

$$= \frac{(100 \times 3) + (250 \times 4) + (150 \times 5)}{100 + 250 + 150} = \frac{2050}{500} = 4.1 \text{ hours.}$$

Now after training the composition of workers is as follows :

Skilled-workers = 100 + 250 = 350
Semi-skilled workers = 150
Unskilled workers = Nil

Average time per worker after training is :

$$= \frac{(350 \times 3) + (150 \times 4)}{350 + 150} = \frac{1050 + 600}{500} = 3.3 \text{ hours.}$$

After 2 years 0.8 hour less would be required.

Note. An assumption has been made that there has been no turnover of workers.

Illustration 28. A Limited Company wants to pay bonus to workers. The bonus is to be paid as under :

Weekly wages (Rs.)					*Bonus (Rs.)*
1300	but not exceeding			1400	1000
1400	"	"	"	1500	1200
1500	"	"	"	1600	1400
1600	"	"	"	1700	1600
1700	"	"	"	1800	1800
1800	"	"	"	1900	2000
1900	"	"	"	2,000	2200
2,000	"	"	"	2,100	2400

Actual wages drawn by the workers is given below :

1325, 1378, 1420, 1620, 1455, 1620 1660, 1680, 1725, 1863, 1832, 1942, 1952,
1800, 2002, 2,028 2,100, 1610, 1625, 1763, 1382, 1540, 1463, 1578, 1723,

How much the company would need to pay by way of bonus ? What shall be the average bonus paid per member of the staff ?

(MBA., Jodhpur Univ., 2005)

Solution. For determining the figure of total and average bonus, first ascertain the number of persons in each wages group.

CALCULATION OF TOTAL AND AVERAGE BONUS

Weekly wages (Rs.)	*Frequency f*	*Bonus (Rs.) X*	*Total Bonus fX*
1300–1400	3	1000	3000
1400–1500	3	1200	3600
1500–1600	2	1400	2800
1600–1700	6	1600	9600
1700–1800	3	1800	5400
1800–1900	3	2000	6000
1900–2,000	2	2200	4400
2,000 and above	3	2400	7200
	$N = 25$		$\Sigma fX = 42,000$

The company would need Rs. 42,000 to pay bonus.

Average bonus per workers = $\frac{42,000}{25}$ = Rs. 1680.

Illustration 29. In 500 small-scale industrial units, the return on investment ranged from 0 to 30 per cent; no units sustaining any loss. 5 per cent of the units had returns ranging from 0 per cent to (and including) 5 per cent, and 15 per cent of the units earned returns exceeding 5 per cent but not exceeding 10 per cent. The median rate of return was 15 per cent and the upper quartile 20 per cent. The uppermost layer of returns exceeding 25 per cent was earned by 50 units.

(*i*) Present the information in the form of a frequency table as follows :

Exceeding 0 per cent but not exceeding 5 per cent
" 5 " " " " 10 "
" 10 " " " " 15 "

and so on.

(*ii*) Find the rate of return around which there is maximum concentration of the units.

Solution. (*i*)

FREQUENCY TABLE

Rate of Return	*Industrial Units*
Exceeding 0 but not exceeding 5 per cent	$500 \times \frac{5}{100} = 25$
" 5 " " 10 "	$500 \times \frac{15}{100} = 75$
" 10 " " 15 "	$250 - 100 = 150$
" 15 " " 20 "	$375 - 250 = 125$
" 20 " " 25 "	$500 - 375 - 50 = 75$
" 25 " " 30 "	50
	Total = 500

(*ii*) For finding out the rate of return around which there is maximum concentration of the units, we will calculate mode. Mode lies in the class 10–15.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 10;\ \Delta_1 = (150 - 75) = 75;\ \Delta_2 = (150 - 125) = 25;\ i = 5$$

$$\text{Mode} = 10 + \frac{75}{75 + 25} \times 5 = 13.75.$$

Hence the maximum concentration is around 13.75 per cent returns.

Illustration 30. The mean monthly salary paid to all employees in a company is Rs. 16000. The mean monthly salaries paid to technical and non-technical employees are Rs. 18000 and Rs. 12000 respectively. Determine the percentage of technical and non-technical employees of the company.

Solution. Let percentage of technical personnel be denoted by X.

Non-technical employees would be $(100 - X)\%$.

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$\bar{X}_{12} = 16,000,\ \bar{X}_1 = 18,000,\ \bar{X}_2 = 12,000$$

$$\text{Let } N_1 = X \ \therefore\ N_2 = (100 - X)$$

Substituting the values

$$16,000 = \frac{18000X + (100 - X)12000}{100}$$

$$1600000 = 18,000X + 1200000 - 12000$$

or $$6000X = 400000$$

$$\bar{X} = \frac{400000}{6000} = 66.67.$$

Hence percentage of technical employees = 66.67 and percentage of non-technical employees = 100 – 66.67 = 33.33.

Illustration 31. A company invests Rs. 1 lakh at 10% annual rate of interest. What will be the total amount after 6 years if the principal is not withdrawn ?

Solution. Applying the compound interest formula :

$$P_n = P_0\left(1+\frac{r}{100}\right)^n$$

$$P_0 = 1,00,000, \quad r = 10, \quad n = 6$$

$$P_n = 1,00,000\left(1+\frac{10}{100}\right)^6.$$

Taki.ag logarithms,

$$\log P_n = \log 1,00,000 + 6\log\left(1+\frac{10}{100}\right)$$

$$= \log 1,00,000 + 6\log 110 - 6\log 100$$

$$= 5 + (6 \times 2.0414) - (6 \times 2) = 5.2484$$

$$P_n = AL\ 5.2484 = 1,77,200.$$

Thus the total amount after 6 years would be Rs. 1,77,200.

Illustration 32. In a certain factory a unit of work is completed by *A* in 10 minutes, by *B* in 15 minutes, by *C* in 12 minutes and by *D* in 20 minutes.

(*i*) What is the average number of units of work completed per minute ?

(*ii*) At this rate how many units will they complete in an 8-hour day ?

Solution. The average number of units of work per minute will be obtained by finding out the harmonic mean

$$H.M. = \frac{4}{\frac{1}{10}+\frac{1}{15}+\frac{1}{12}+\frac{1}{20}} = \frac{4\times 120}{36} = \frac{40}{3}$$

Hence $\frac{40}{3}$ units per minute is the average rate.

The average units per minute is $\frac{1}{\frac{40}{3}} = \frac{3}{40}$

(*b*) In 8 hours, *i.e.*, 480 minutes, the total number of units produced will be $480 \times \frac{3}{40} \times 4 = 144$ by all the four workers combined.

Illustration 33. A machine was purchased for Rs. 10 lakh in 2006. Depreciation on the diminishing balance was charged @ 40% in the first year, 25% in the second year and 10% per annum during the next three years. What is the average depreciation charge during the whole period?

Solution. Since we are interested in finding out the average rate of depreciation, geometric mean will be the most appropriate average. The cost of machine can be ignored as it is immaterial in the rate calculation.

DETERMINING AVERAGE RATE OF DEPRECIATION

Year	*Diminishing value (for a value of Rs. 100)* X	*Log X*
2006	100 – 40 = 60	1.7782
2007	100 – 25 = 75	1.8751
2008	100 – 10 = 90	1.9542
2009	100 – 10 = 90	1.9542
2010	100 – 10 = 90	1.9542
		$\Sigma \log X = 9.5159$

$$\text{G.M.} = \text{AL}\left(\frac{\Sigma \log X}{N}\right) = \text{AL}\left(\frac{9.5159}{5}\right) = \text{AL } 1.9032 = 80.$$

The diminishing value being Rs. 80, the depreciation will be 100 – 80 = 20%.

Illustration 34. The following is the age distribution of 1,000 person working in a large industrial house :

Age group	*No. of persons*	*Age group*	*No. of persons*
20–25	30	45–50	105
25–30	160	50–55	70
30–35	210	55–60	60
35–40	180	60–65	40
40–45	145		

Due to continuous losses, it is desired to bring down the strength to 30% of the present number according to the following scheme :

(*i*) To retrench the first 15% from the lower group.

(*ii*) To absorb the next 45% in other branches.

(*iii*) To make 10% from the highest age group retire permanently, if necessary.

Calculate the age limits of the persons retained and those to be transferred to other departments. Also find the average age of those retained. *(MBA,DU, 2005)*

Solution. The first 15% to be retrenched are from the lower group ; hence their total number comes to $\frac{15 \times 1{,}000}{100} = 150$. 30 belong to 20 – 25 age group and rest, *i.e.*, (150–30), *i.e.*, 120 belong to 25–30 age group. The next 45% are to be absorbed in other branches. They are $1000 \times \frac{45}{100} = 450$. They belong to the following age groups :

Age groups	*No. of persons*
25–30	40
30–35	210
35–40	180
40–45	20

Those to retire are 10% and belong to the highest age group. Their number comes to $1{,}000 \times \frac{10}{100} = 100$ and their age groups are :

Age groups	*No. of persons*
60–65	40
55–60	60

AVERAGE AGE OF THOSE RETAINED

Age groups	*m.p.* X	*No. of persons* f	$(X - 47.5)/5$ d	fd
40–45	42.5	125	–1	–125
45–50	47.5	105	0	0
50–55	52.5	70	+1	+70
		$N = 300$		$\Sigma fd = -55$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 47.5 - \frac{55}{300} \times 5 = 47.5 - 0.92 = 46.58.$$

Hence the average age of those retained is 47 years approximately.

Illustration 35. The median and mode of the following wage distribution are Rs. 33.5 and Rs. 34 respectively. However, three frequencies are missing. Determine their values.

Wages (in hundred Rs.)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	Total
Frequencies	4	16	?	?	?	6	4	230

Solution. Let the missing frequencies be f_0, f_1 and f_2 corresponding to classes 20–30, 30–40 and 40–50 respectively. Since median and mode are 33.5 and 34, they lie in the class 30–40. The frequency of this class is f_1.

DETERMINING MISSING VALUES

Wages (in hundred Rs.)	*Frequency*	*Cum. frequency*
0–10	4	4
10–20	16	20
20–30	?	$20 + f_0$
30–40	?	$20 + f_0 + f_1$
40–50	?	$20 + f_0 + f_1 + f_2$
50–60	6	226
60–70	4	230
	$N = 230$	

$$f_0 + f_1 + f_2 = 230 - (4 + 16 + 6 + 4) = 200$$

$$f_2 = 200 - (f_0 + f_1) = 200 - f_0 - f_1$$

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$34 = 30 + \frac{f_1 - f_0}{2f_1 - f_0 - (200 - f_0 - f_1)} \times 10$$

$$\frac{4}{10} = \frac{f_1 - f_0}{2f_1 - f_0 - 200 + f_0 + f_1}$$

$$10f_1 - 10f_0 = 4\,(3f_1 - 200) \text{ or } 10f_1 - 10f_0 = 12f_1 - 800$$

$$-2f_1 - 10f_0 = -800 \text{ or } f_1 + 5f_0 = 400. \quad \ldots(i)$$

$$\text{Med.} = L + \frac{N/2 - \text{p.c.f.}}{f} \times i$$

$$33.5 = 30 + \frac{115 - 20 - f_0}{f_1} \times 10$$

$$3.5f_1 = 950 - 10f_0$$

$$7f_1 = 1{,}900 - 20f_0$$

$$7f_1 + 20f_0 = 1{,}900 \quad \ldots(ii)$$

From Eqns. (*i*) and (*ii*),

$$f_1 + 5f_0 = 400$$

$$7f_1 + 20f_0 = 1{,}900$$

Multiplying Eqn. (*i*) by 4,

$$4f_1 + 20f_0 = 1{,}600$$

$$7f_1 + 20f_0 = 1{,}900$$

$$-3f_1 = -300 \text{ or } f_1 = 100$$

Substituting the value of f_1 in Eqn. (*i*),

$$100 + 5f_0 = 400$$

$$5f_0 = 300 \text{ or } f_0 = 60$$

Since

$$f_0 + f_1 + f_2 = 200$$

$$f_2 = 200 - 100 - 60 = 40$$

Hence the missing frequencies are

$$f_0 = 60, f_1 = 100, f_2 = 40.$$

Illustration 36. Calculate the arithmetic mean and the median of the frequency distribution given below. Hence calculate the mode using the empirical relation between the three :

Height (in cms.)	*No. of students*	*Height (in cms.)*	*No. of students*
130–134	5	150–154	17
135–139	15	155–159	10
140–144	28	160–164	1
145–149	24		

Solution. **CALCULATION OF ARITHMETIC MEAN AND MEDIAN**

Height (in cms.)	*m.p.* X	f	$(X-147)/5$ d	fd	*c.f.*
129.5–134.5	132	5	–3	–15	5
134.5–139.5	137	15	–2	–30	20
139.5–144.5	142	28	–1	–28	48
144.5–149.5	147	24	0	0	72
149.5–154.5	152	17	+1	+17	89
154.5–159.5	157	10	+2	+20	99
159.5–164.5	162	1	+3	+3	100
		$N = 100$		$\Sigma fd = -33$	

Mean : Mean $\bar{X} = A + \frac{\Sigma fd}{N} \times i = 147 - \frac{33}{100} \times 5 = 147 - 1.65 = 145.35$

Median : Median = Size of $\frac{N}{2} = \frac{100}{2}$ = 50th observation.

Median lies in the class 144.5 – 149.5.

$$\text{Median} = L + \frac{N/2 - p.c.f.}{f} \times i = 144.5 + \frac{50-48}{24} \times 5 = 144.5 + .417 = 144.917.$$

Mode : Mode = 3 Median – 2 Mean = (3 × 144.917) – (2 × 145.35) = 144.051.

Illustration 37. Income of employees in an industrial concern are given below. The total income of the 10 employees in the class over Rs. 25000 is Rs. 3,00,000. Compute the mean income. Every employee belonging to the top 25% of the earners is required to pay 5% of his income to workers' relief fund. Estimate the contribution to this fund.

Income (Rs.)	*Employers*	*Income (Rs.)*	*Employers*
Below 5000*	90	15000–20000	80
5000–10000	150	20000–25000	70
10000–15000	100	25000 and over	10

Solution. **COMPUTATION OF MEAN**

Income (Rs.)	*m.p.(X)*	f	fX
0–5000	2500	90	2,25,000
5000–10,000	7500	150	11,25,000
10,000–15,000	12500	100	12,50,000
15,000–20,000	17500	80	14,00,000
20,000–25,000	22500	70	15,75,000
25,000 and over	30000 (given)	10	3,00,000
		$N = 500$	$\Sigma fX = 58,75,000$

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{587500}{500} = \text{Rs. } 11750.$$

Number of employees belonging to the top 25% of the earners is $\frac{25}{100} \times 500 = 125$ and the distribution of these top earners as obvious from the above table is as follows :

Distribution of top 25% earners

* Since class intervals are equal, below 5000 should mean 0–5000

Income (Rs.)	*Frequency*
25000 and over	10
20000–25000	70
15000–20000	45

45 persons are to be taken in the last class, *i.e.*, 15000–20000 with the highest income level starting from 20000 and below. Under the assumption that the frequencies are equally distributed throughout the class, the calculations would be as follows :

80 persons have income in the range 15000–20000 = Rs. 5000

$\therefore$ 45 persons have income in the range = $\frac{5000}{80} \times 45 = 2812.5$ or 2812

Since we are interested in the top 45 earners in the income group 15000–20000, their salaries will range from (20000–2812) to 20000, *i.e.*, 17188 to 20000.

The distribution of top 125 persons is as follows :

Income (Rs.)	*m.p. (X)*	*f*	*Total Income fX*
25000 and over	—	10	300000 (given)
20000–25000	22500	70	1575000
17188–20000	18594	45	836730
		$N = 125$	$\Sigma fX = 2711730$

Hence the total income of the top 25% of earners is Rs. 27,11,730

5% Contribution to the fund = 0.05×2711730 = Rs. 135586.5

Illustration 38. Calculate the median and mode for the distribution of the weights of 150 students from the data given below :

Weight (in kg) :	30–40	40–50	50–60	60–70	70–80	80–90
Frequency :	18	37	45	27	15	8

Solution.

CALCULATION OF MEDIAN AND MODE

Weight (kg)	*f*	*c.f.*
30–40	18	18
40–50	37	55
50–60	45	100
60–70	27	127
70–80	15	142
80–90	8	150
	$N = 150$	

Median = Size of $\frac{N}{2}$ th observation = $\frac{150}{2}$ = 75th observation

Median lies in the class 50–60.

Median = $L + \frac{N/2 - p.c.f.}{f} \times i = 50 + \frac{75-55}{45} \times 10 = 50 + 4.444 = 54.44$

Mode : Since highest frequency is 45, mode lies in the class 50–60.

Mode = $L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 50 + \frac{8}{8+18} \times 10 = 50 + 3.08 = 53.08.$

Illustration 39. Following distribution gives the pattern of overtime work per week done by 100 employees of a company. Calculate median, first quartile, and 7th decile.

Overtime hours :	10–15	15–20	20–25	25–30	30–35	35–40
No. of employees :	11	20	35	20	8	6

(MBA, Kurukshetra Univ., 2

Solution.

CALCULATION OF MEDIAN, Q_1 AND D_7

Overtime (hrs.)	*f*	*c.f.*
10–15	11	11
15–20	20	31
20–25	35	66
25–30	20	86
30–35	8	94
35–40	6	100
	$N = 100$	

Med. = Size of $\frac{N}{2}$ th observation = $\frac{100}{2}$ = 50th observation

Median lies in the class 20–25.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i = 20 + \frac{50-31}{35} \times 5 = 20 + 2.714 = 22.714$$

$$Q_1 = \text{size of } \frac{N}{4} \text{ th observation} = \frac{100}{4} = 25\text{th observation}$$

Q_1 lies in the class 15–20.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 15 + \frac{25-11}{20} \times 5 = 15 + 3.5 = 18.5$$

$$D_7 = \text{Size of } \frac{7N}{10} \text{ th observation} = \frac{7 \times 100}{10} = 70\text{th observation}$$

D_7 lies in the class 25–30.

$$D_7 = L + \frac{7N/10 - p.c.f.}{f} \times 5 = 25 + \frac{70-66}{20} \times 5 = 25 + 1 = 26.$$

Illustration 40. In an examination of 675 candidates the examiner supplied the following information :

Marks obtained	*No. of candidates*	*Marks obtained*	*No. of candidates*
Less than 10%	7	Less than 50%	381
Less than 20%	39	Less than 60%	545
Less than 30%	95	Less than 70%	631
Less than 40%	201	Less than 80%	675

Calculate the mode and median of the percentage marks obtained. *(MBA, Rohilkhand Univ., 2002)*

Solution. **CALCULATION OF MEDIAN AND MODE**

Marks (in %)	*No. of candidates* (f)	*m.p.* Xs	$(X-45)/10$ d	fd	*c.f.*
0–10	7	5	–4	–28	7
10–20	32	15	–3	–96	39
20–30	56	25	–2	–112	95
30–40	106	35	–1	–106	201
40–50	180	45	0	0	381
50–60	164	55	+1	+164	545
60–70	86	65	+2	+172	631
70–80	44	75	+3	+132	675

Median : Med. = Size of $\frac{N}{2}$ th observation = $\frac{675}{2}$ = 337.5th observation.

Median lies in the class 40–50.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i$$

$L = 40$, $N/2 = 337.5$, $p.c.f. = 201$, $f = 180$, $i = 10$

$$\text{Med.} = 40 + \frac{337.5-201}{180} \times 10$$

$$= 40 + 7.58 = 47.58$$

Mode : By inspection mode lies in the class 40–50.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$L = 40, \Delta_1 = 180 - 106 = 74, \Delta_2 = 180 - 164 = 16, i = 10.$

$$\text{Mode} = 40 + \frac{74}{74+16} \times 10 = 40 + 8.22 = 48.22.$$

Illustration 41. The following is the average amount of dollars each major airline spends per passenger on food

American	7.41	Continental	2.77
United	7.24	US Air	2.68
Northwest	5.15	American west	2.00
TWA	5.09	Southwest	0.14
Delta	4.61		

What are the mean and median cost per passenger? Which would be the better figure to use for a new airline in developing its business plan? *(MBA, Delhi Univ., 2003)*

Solution. Calculation of $\bar{X}$ and median.

$$\bar{X} = \frac{7.41 + 7.24 + 5.15 + 5.09 + 4.61 + 2.77 + 2.68 + 2.00 + 0.14}{9} = \frac{37.09}{9}$$

$= \$ 4.12.$

Median. Arranging the given data in ascending order.

0.14 2.00 2.68 2.77 4.61 5.09 5.15 7.24 7.41

$$\text{Med} = \text{Size of } \frac{N+1}{2} \text{ th observation}$$

$$= \frac{9+1}{2} = 5\text{th observation}$$

Size of 5th observation is 4.61. Hence median = $ 4.61.

Median would be a better choice for new airlines in developing its business plan as median is not affected by extreme observations.

Illustration 42. Consider the following distribution :

X	0–10	10–20	20–30	30–40	40–50
f	12	18	20	25	23

Compute mean median and mode. *(MBA, GGS.I.P. Univ., 2009)*

Solution. **CALCULATION OF MEAN, MEDIAN AND MODE**

X	f	$m.p.$ X	$(X-25)/10$ d	fd	$c.f.$
0–10	12	5	–2	–24	12
10–20	18	15	–1	–18	30
20–30	20	25	0	0	50
30–40	25	35	+1	+25	75
40–50	23	45	+2	+46	98
	$N = 98$			$\Sigma fd = 29$	

Mean : Mean $\bar{X} = A + \frac{\Sigma fd}{N} \times i = 25 + \frac{29}{98} \times 10 = 25 + 2.96 = 27.96$

Median : Med. = Size of $\frac{N}{2}$ th item $= \frac{98}{2} = 49$th item

Median lies in the class 20–30.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i$$

$L = 20, N/2 = 49, p.c.f. = 30, f = 20, i = 10$

$$= 20 + \frac{49-30}{20} \times 10 = 20 + 9.5 = 29.5$$

Mode : Mode lies in the class 30–40.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$L = 30,\ \Delta_1 = (25 - 20) = 5,\ \Delta_2 = (25 - 23) = 2,\ i = 10.$

$$\text{Mode} = 30 + \frac{5}{5+2} \times 10 = 30 + \frac{50}{7} = 30 + 7.14 = 37.14.$$

PROBLEMS

Answer following questions, each question carries **one** mark.

(*i*) What is arithmetic mean? (*MBA, Madurai Kamaraj Univ., 2001*)
(*ii*) What is meant by mode?
(*iii*) What is the empirical formula linking mean, median and mode? (*MBA, Madurai Kamaraj Univ., 2004*)
(*iv*) Give formula for geometric mean and harmonic mean in case of a continuous frequency distribution.
(*v*) When mode is ill-defined?
(*vi*) What are quartiles and percentiles?
(*vii*) Is sum of deviations from arithmetic mean always zero?
(*viii*) What is weighted mean?
(*ix*) Is harmonic mean reciprocal of arithmetic mean?
(*x*) Why arithmetic mean is most affected by extreme observations?
(*xi*) When is mode useful over other averages?

Answer the following questions, each question carries **four** marks.

(*i*) Define average. Write down the properties of an average. (*MA Eco., M.K. Univ., 2003*)
(*ii*) What are the uses of geometric mean and harmonic mean?
(*iii*) What is combined mean? Explain with the help of an example.
(*iv*) Distinguish between median, quartiles, deciles and percentiles.
(*v*) What is the arithmetic means of first n natural number, 1, 2 ..... n ?

What are the various measures of central tendency? Why are they called measures of central tendency? (*MBA, UP. Tech. Univ., 2004*)

Give a brief description of different measures of central tendency. Why is arithmetic mean so popular?

(*a*) Is it necessarily true that being above average indicates that someone is superior? Explain.
(*b*) What are quartiles of a distribution? Explain their uses.

(*a*) Define arithmetic mean and median and discuss their merits and demerits as measures of central tendency.
(*b*) How would you account for the predominant choice of arithmetic mean as a measure of central tendency? Under what circumstances would it be appropriate to use mode, median, geometric mean and harmonic mean. (*MBA, KU, 2002; MBA, Delhi Univ., 2006*)

Comment on the following :

(*i*) If first and third quartiles are 20 and 40 respectively the median will be 30.
(*ii*) If daily wages paid to men and women employed in a factory are Rs. 100 and Rs. 90, the average wage per worker would be Rs. 90.
(*iii*) A man claims that his average bank balance during the year is Rs. 3700. The bank, on the other hand, claims that he overdrew his account at least 10 times during the year and as such his claim is false.
(*iv*) The increase in the price of commodity x is 20%. Then the price decreased 25% and again increased 15%. The resultant increase in the price is 10%.
(*v*) The mode of a distribution cannot be less than the arithmetic mean.
(*vi*) If Q_1, Q_2, Q_3, be respectively the lower quartile, the median and the upper quartile of a distribution, then $Q_2 - Q_1 = Q_3 - Q_2$.
(*vii*) Arithmetic mean is the best measure of central tendency.

What is a statistical average? What are the desirable properties for an average to possess? Mention different types of averages and state why the arithmetic mean is the most commonly used amongst them.

(*a*) What are the essential requisites of a good measure of central tendency? Compare and contrast the commonly employed measure in terms of these requisites.
(*b*) Prove that the arithmetic mean of two positive numbers a and b is at least as large as their geometric mean.

(*a*) What are the properties of a good average?
(*b*) In each of the following cases, explain whether the description applies to mean, median or both :
(*i*) Can be calculated from a frequency distribution with open-end classes?
(*ii*) The values of all observations are taken into consideration in the calculation.
(*iii*) The values of extreme observations do not influence the average.

(*a*) Under what circumstances would it be appropriate to use Arithmetic mean, Median or Mode? Discuss.

(*b*) Explain the properties of a good average. In the light of these properties which average do you think is the best and why? (*MBA, Jodhpur Univ.*

11. (*a*) Give a brief note of the measures of central tendency together with their merits and demerits. Which is the best measure of central tendency and why? (*MBA, Osmania Univ.*
(*b*) "Every average has its own peculiar characteristics. It is difficult to say which average is the best." Comment [*MBA, HPU*

12. For the following frequency table calculate mean, median and mode :

Weekly rent (in Rs.)	*No. of persons paying the rent*	*Weekly rent (in Rs.)*	*No. of persons paying the rent*
200–400	6	1,200–1,400	15
400–600	9	1,400–1,600	10
600–800	11	1,600–1,800	8
800–1,000	14	1,800–2,000	7
1,000–1,200	20		

[Mean = Med. = 1,100; Mode = 1,109.41]

13. Calculate the simple and weighted arithmetic mean price per bag of 20 kg. of coal purchased by an industry for the half year. Account for difference between the two.

Month	*Price per bag (Rs.)*	*Bag purchased*	*Month*	*Price per bag (Rs.)*	*Bag purchased*
Jan.	42.05	25	April	52.00	52
Feb.	51.25	30	May	44.25	10
Mar.	50.00	40	June	54.00	45

[$\bar{X} = 48.93$; $\bar{X}_w = 50.31$]

14. (*a*) Explain clearly the concepts of Geometric mean and Harmonic mean. Point out some of the business applications of these concepts.
(*b*) Calculate the geometric mean of the following price relatives :

Commodity	*Price Relative*	*Commodity*	*Price Relative*
Wheat	237	Sugar	124
Rice	198	Salt	107
Pulses	156	Oils	196

[G.M. = 163.4]

15. The following table gives the distribution of 100 accidents during seven days of the week of a given month. During the particular month there were 5 Mondays, Tuesdays and Wednesdays, and only four each of the other days. Calculate the number of accidents per day.

Days	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
No. of accidents	26	16	12	10	8	10	18

[14.13]

16. At harvesting time a farmer employed 10 men, 24 women and 16 boys to lift potatoes. A woman's work was three quarters as effective as that of a man, while a boy's work was only half. Find the daily wage bill if a man rate was Rs. 80 a day and the rates for the women and boys in proportion to their effectiveness. Calculate the average daily rate for all the workers.

17. The following table gives the daily wages (in rupees) in a certain commercial organisation :

Daily wages (Rs.)	*30–32*	*32–34*	*34–36*	*36–38*	*38–40*
No. of Workers	2	9	25	30	49
Daily wages (Rs.)	40–42	42–44	44–46	46–48	48–50
No. of Workers	62	39	20	11	3

Calculate from the above data :
(*i*) the median and the third quartile wages; and
(*ii*) the number of wage-earners receiving between Rs. 37 and Rs. 45.
[(*i*) Med. 40.32; $Q_3 = 42.54$, (*ii*) 175]

18. Six types of workers are employed in each of two workshops but at different rates of wages as follows :

Types of workers	*Workshop A Daily wages per worker*	*No. of workers*	*Wrokshop B Daily wages per worker*	*No. of workers*
Mechanic	92.50	2	93.00	11
Fitter	93.50	14	93.00	50
Electrician	94.00	20	94.25	8
Carpenter	93.00	7	93.50	10
Smith	93.00	6	93.50	10
Clerk	92.00	1	95.00	2

In which of the two workshops is the average rate of wages per worker higher and by how much?

19. (*a*) A motor car covered a distance of 50 km four times. The first time at 50 km p.h., the second at 20 km p.h., the third at 40 km p.h. and the fourth at 25 km p.h. Calculate the average speed and explain the choice of the average.
[29.63]

(*b*) A man gets three annual increases in salary. At the end of the first year he gets an increase of 4%, at the end of second year an increase of 6% on his salary as it was at the end of the first year, and at the end of the third year an increase of 9% on his salary as it was at the end of the second year. What is the average percentage increase?
[6.7]

(*c*) A machine is assumed to depreciate 44 per cent in value in the first year, 25% in the second year and 10% per annum for the next three years, each percentage being calculated on the diminishing value. What is the average percentage depreciation for the five years? *(MBA. Vikram Univ., 2001)*
[21.07%]

20. (*a*) Mr. *A* spends Rs. 1000 for apples costing Rs. 25 per kilogram and another Rs. 1000 for apples costing Rs. 20 per kilogram. What is the average price of apples per kilogram? *(MBA, Vikram Univ., 2005)*
[22.22]

(*b*) Three men take 12, 8, 6 hours respectively to husk an acre of corn. Determine the average number of hours to husk an acre.
[$\bar{X}$ = 8.67]

21. (*a*) If oranges for one rupee are bought at 10 paise each and for another rupee at 5 paise each, the average price would be $6\frac{2}{3}$ paise and not $7\frac{1}{2}$ paise. Explain and verify.

(*b*) You take a trip which entails travelling 900 kilometres by train at an average speed of 60 kilometres per hour, 3,000 kilometres by boat at an average of 25 kilometres per hour, 400 kilometres by plane at 350 kilometres per hour and finally 15 kilometres by taxi at 25 kilometres per hour. What is your average speed for the entire distance?
[31.6]

(*c*) A certain store made weekly profits of Rs. 5,000, Rs. 10,000 and Rs. 80,000 in 2008, 2009 and 2010 respectively. Determine the average rate of growth of this store's profits.
[266.6]

22. The following distribution represents the number of minutes spent by a group of teenagers in going to movies. What is the median?

Minutes/Week	*Number of teenagers*	*Minutes/Week*	*Number of teenagers*
0–99	27	400–499	58
100–199	32	500–599	38
200–299	65	600 and more	9
300–399	78		

[Med. = 333.6]

23. An investor buys Rs. 1,200 worth of shares in a company each month. During the first five months he bought the shares at a price of Rs. 10, Rs. 12, Rs. 15, Rs. 20 and Rs. 24 per share. After 5 months what is the average price paid for the shares by him?
[14.63]

24. The value of a machine decreases at a constant rate from the cost price of Rs. 10,000 to scrap value of Rs. 1,000 in ten years. Find the annual rate of decrease and the value of the machine at the end of one, two and three years.
[25.89]

25. An incomplete distribution is given below :

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

You are told that the median value is 46. Using the median formula, fill up the missing frequencies and calculate the arithmetic mean of the completed table.
[Freq. corresponding to 30 – 40 is 33.5; and corresponding to 50 – 60 is 45.5 ; and $\bar{X} = 45.8\dot{7}$]

26. The production of butter fat during 7 consecutive days was recorded for 300 cows. Calculate the average fat content.

Butter fat (lb) :	10–,	11—,	12—,	13—,	14—,	15—,	16—,	17—,
No. of cows :	8	25	50	75	60	48	22	12

[14 lb]

27. The number of telephone calls received in 293 successive one-minute intervals at an exchange are shown in the following table :

No. of calls :	0	1	2	3	4	5	6	7	8
Frequency :	10	28	35	45	65	52	32	12	14

Calculate mean and modal number of calls.
[3.89 ; 4]

8. Given below is the frequency distribution of the marks obtained by 90 students. Compute the arithmetic mean, median and mode :

Marks	*No. of Students*	*Marks*	*No. of Students*
20–29	2	60–69	18
30–39	12	70–79	10
40–49	15	80–89	9
50–59	20	90–99	4

(58.5, 57.5, 56.64)

29. The geometric mean of 10 observations on a certain variable was calculated to be 16.2. It was later discovered that one of the observations was wrongly recorded at 10.9 when in fact it was 21.9. Apply appropriate correction and calculate the correct geometric mean.

[17.08]

30. A recent college graduate was hired by a large manufacturing corporation and placed in their management training programme. As part of her training, she was assigned to five different departments of the corporation for various periods of time. At the end of the training period in each department, the supervisor graded her performance on a scale from zero to ten. At the end of the training programme, the training director computed an overall mean score based on the following consideration :

The marketing and production phases of her training were assumed to be of equal importance. Both of these were considered to be three times as important as the purchasing and financial phases. The accounting training was twice as important as the latter two.

If the supervisor's ratings were as follows, compute an appropriate mean score :

Department	*Score*
Finance	4
Marketing	7
Production	8
Purchasing	6
Accounting	9

31. Atul gets a pocket money allowance of Rs. 120 per month. Thinking that this was rather less, he asked his friends about their allowances and obtained the following data which includes his allowance also (Amounts in Rs.) :

112, 118, 110, 115, 125, 120, 120, 122, 115, 110, 110, 115, 113, 120,
118, 110, 115, 110, 117, 118, 115, 112, 115, 110, 115, 110, 112, 118,
120, 125, 118.

He presented these data to his father and asked for an increase in his allowances as he was getting less than average amount. His father, a statistician, countered pointing out that Atul's allowance was actually more than the average amount.

Reconcile these statements.

Would Atul or his friends getting less than the 'average' have no more cause for complaint if their allowances were increased to the 'average' amount? Give reasons for your answer.

32. 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 percentage of the workers who earned between Rs. 1750 and Rs. 2250,
(*f*) The percentage who earned more than Rs. 2500 per week, and
(*g*) The percentage who earned less than Rs. 2000 per week.

Weekly wage (Rs.)	*No. of workers*	*Weekly wage (Rs.)*	*No. of workers*
1200–1400	8	2200–2400	32
1400–1600	12	2400–2600	18
1600–1800	20	2600–2800	7
1800–2000	30	2800–3000	6
2000–2200	40	3000–3200	4

33. From the following distribution of travel time to work of a firm's employees, find the modal travel time :

Travel time (in minutes)	*Frequency*
Less than 80	218
Less than 70	215
Less than 60	195
Less than 50	156
Less than 40	85
Less than 30	50
Less than 20	18
Less than 10	2

[illegible] the following incomplete frequency distribution. It is known that the total frequency is 1,000 and that the median is 413.11. Estimate by calculation the missing frequencies and find the value of the mode.

Sales (Rs. lakhs)	*No. of companies*	*Sales (Rs. lakhs)*	*No. of companies*
300–325	5	400–425	326
325–350	17	425–450	7
350–375	80	450–475	88
375–400	?	475–500	9

[127, 248; mode = 413.98]

35. (*a*) In a certain office a letter is typed by *A* in 4 minutes. The same letter is typed by *B*, *C* and *D* in 5, 6, 10 minutes respectively. What is the average time taken in completing one letter? How many letters do you expect to be typed in one day comprising 9 working hours?

[H.M. = 5.58 minutes per letter.
Letters typed in 8 hours (480 minutes) = 86]

(*b*) For an income distribution of a group of men, 20 p.c. of men have income below Rs. 3500, 35 p.c. below Rs. 7500, 60 p.c. below Rs. 17500 and 80 p.c. below Rs. 25000; the first and third quartile are Rs. 5500 and Rs. 20000. Put the above information in cumulative frequency distribution and find the median.

36. The rate of a certain commodity in the first week of January 2008 was 0.4 kg per rupee; it was 0.6 kg per rupee in the second week and 0.5 kg per rupee in the third week. Therefore, it is correct to say that the average price was 0.5 kg per rupee. Verify.

37. Below given is the frequency distribution of weekly wages of 100 workers in a factory :

Weekly wages (Rs.)	*No. of workers*	*Weekly wages (Rs.)*	*No. of workers*
1120–1124	3	1145–1149	10
1125–1129	5	1150–1154	8
1130–1134	12	1155–1159	5
1135–1139	23	1160–1164	3
1140–1144	31		

Draw the ogive for the distribution and use it to determine the median wage of a worker and verify the result by the formula. How many workers earned weekly wages between Rs. 1132 and Rs. 1153?

38. Find the missing frequencies in the following distribution if *N* is 100 and median 30 :

Marks :	0–10	10–20	20–30	30–40	40–50	50–60	
No. of students :	10	15	?	30	10	8	*(MBA, M.D. Univ., 1999)*

39. Draw an ogive for the following distribution. Read the median from the graph and verify your result by the mathematical formula. Also obtain the limits of income of central 50% of the employees.

Weekly Income (Rs.)	*No. of employees*	*Weekly Income (Rs.)*	*No. of employees*
Below 550	6	700–750	16
500–600	10	750–800	12
600–650	22	Above 800	15
650–700	30		*(MBA, Delhi Univ., 1999)*

[Med. = 679.2; 626.7 to 747.7]

40. Following is the cumulative frequency distribution of preferred length of kitchen slabs obtained from the preference study on 50 housewives.

Length (in metres) more than	*Number of housewives*	*Length (in metres) more than*	*Number of housewives*
1.0	50	2.5	42
1.5	46	3.0	10
2.0	40	3.5	3

A manufacturer has to take decision on what length of slabs to manufacture. What length would you recommend and why?

41. Following are the data for marks obtained by students in a paper. The top 20% students will qualify for a prize. What is the lower limit of marks above which the student will get the prize?

Marks	*No. of students*	*Marks*	*No. of students*
0–10	5	50–60	10
10–20	7	60–70	4
20–30	8	70–80	4
30–40	10	80–90	2
40–50	10		

42. A factory pays workers on piece rate basis and also a bonus to each worker on the basis of individual output in each quarter. The rate of bonus payable is as follows :

Output (in units)	*Bonus (in rupees)*	*Output (in units)*	*Bonus (in rupees)*
70–74	400	90–94	700
75–79	450	95–99	800
80–84	500	100–104	1000
85–89	600		

The individual output of a batch of 50 workers is given below :

94	83	78	76	88	86	93	80	91	82
89	97	92	84	932	80	85	83	98	103
87	88	88	81	95	86	99	81	87	90
84	97	80	75	93	101	82	82	89	72
85	83	75	72	83	98	77	87	71	80

By suitable classification you are required to find :

(*i*) Average bonus per worker for the quarter.

(*ii*) Average output per worker.

[(*i*) 90.03 (*ii*) 86.1]

43. An individual purchases three qualities of ball-pens. The relevant data are given below :

Quality	*Price per ball-pen (Rs.)*	*Money spent (Rs.)*
A	10.00	500
B	10.50	300
C	20.00	200

Calculate an average price per ball-pen. (*MBA, Kurukshetra Univ., 2002*)

44. A number of particular articles have been classified according to their weights. After drying for two weeks the same articles have been weighted and similarly classified. It is known that the median weight in the first weighing it was 17.35. Some frequencies in the first weighing (a and b) and second weighing (x and y) are missing. It is known that $a = 1/3x$ and $b = 1/2y$. Find out the value of a, b, x and y.

	1st Weighing	*IInd Weighing*
0–5	a	x
5–10	b	y
10–15	11	40
15–20	52	50
20–25	75	30
25–30	22	28

[$a = 3$, $b = 6$, $x = 9$, $y = 12$]

45. Describe the method of constructing ogive. How would you determine median from it? Draw ogive and find median from the following data :

Marks :	0–15	15–30	30–45	45–60	60–75	
No. of Students :	2	15	30	9	4	(*M. Com., AMU, 2001*)

Calculate the median and quartiles for the following data :

Class-Interval	*Frequency*	*Class-Interval*	*Frequency*
0–50	20	150-200	30
50-100	60	200-250	24
100-150	50	250-300	16

46. Calculate the mean and median for the following data :

Central wage (in Rs.) :	15	20	25	30	35	40	45
No. of wage earners :	3	25	19	16	4	5	6

(*MBA, Madurai Kamaraj Univ., 2007*)

*****

5

Measures of Variation

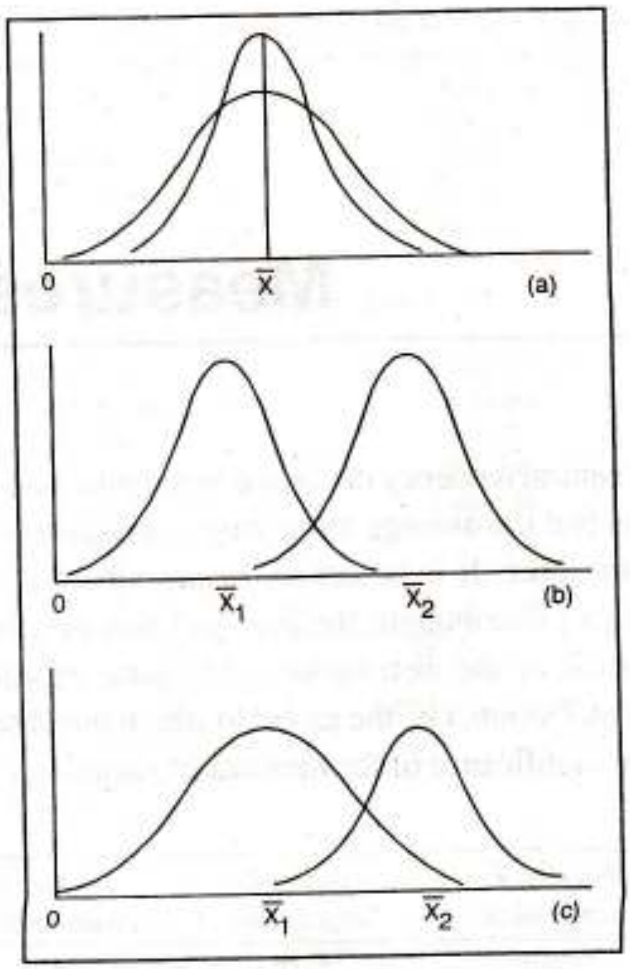
INTRODUCTION

The various measures of central tendency discussed in the previous chapter give us one single value that represents the entire data. But the average alone cannot adequately describe a set of observations, unless all the observations are alike. It is necessary to describe the *variability or dispersion of the observations. Also in* two or more distributions the average value may be the same but still there can be wide disparities in the formation of the distributions. Measures of variation help us in studying the important characteristic of a distribution, *i.e., the extent to which the observations vary from one another from some average value. The significance* of the measure of variation can best be appreciated from the following example :

	Factory A wages (Rs.)	*Factory B wages (Rs.)*	*Factory C wages (Rs.)*
	2300	2310	2380
	2300	2300	2210
	2300	2304	2220
	2300	2306	2200
	2300	2280	2490
Total :	11,500	11,500	11,500
$\bar{X}$:	2,300	2,300	2,300

The above data pertains to five workers each in three different factories. Since the average wage is the same in all factories, one is likely to conclude that the factories are alike in their wage structure, but a close examination shall reveal that the wage distribution in the three factories differs widely from one another. In factory *A,* each and every worker is perfectly represented by the arithmetic mean, *i.e.,* average wage or, in other words, none of the workers of factory *A* deviates from the arithmetic mean and hence there is no variation. In factory *B*, only one worker is perfectly represented by the arithmetic mean, the other workers vary from the mean but the variation is very small as compared to the workers of factory *C*. In factory *C*, the mean does not represent the workers as the individual wage figures differ widely from the mean. Thus we find there is no variation in the wages of workers in factory *A*, there is very little variation in factory *B* but the wages of workers of factory *C* differ most widely. For the student of social sciences, the mean wage is not so important as to know how these wages are distributed. Are there a large number receiving the mean wage or are there a few with enormous wages and millions with wages far below the mean? The following three diagrams (given on next page) represent frequency distribution with some of the characteristics we wish to emphasise :

The two curves in diagram (*a*) represent two distributions with the same mean $\bar{X}$, but with different variations. The two curves in (*b*) represent two distributions with the same variations but with unequal means, $\bar{X}_1$, and $\bar{X}_2$, Finally, (*c*) represents two distributions with unequal means and unequal variations.



The measures of central tendency are, therefore, insufficient. They must be supported and supplemented with other measures. In this chapter, we shall be especially concerned with the measures of variation (or spread, or dispersion). A measure of variation is designed to state the extent to which the individual measures* differ on an average from the mean. In measuring variation, we shall be interested in the amount of the variation or its *degree* but not in the *direction.*** For example, a measure of 6 centimetres below the mean has just as much variation as a measure of 6 centimetres above the mean.

Significance of Measuring Variation

Measures of variation are needed for four basic purposes :

(*i*) To determine the reliability of an average;

(*ii*) To serve as a basis for the control of the variability;

(*iii*) To compare two or more series with regard to their variability; and

(*iv*) To facilitate the use of other statistical measures.

A brief explanation of these points is given below :

(*i*) Measures of variation point out as to how far an average is representative of the entire data. When variation is small, the average is a typical value in the sense that it closely represents the individual value and it is reliable in the sense that it is good estimate of the average in the corresponding universe. On the other hand, when variation is large, the average is not so typical, and unless the sample is very large, the average may be quite unreliable.

(*ii*) Another purpose of measuring variation is to determine nature and cause of variation in order to control the variation itself. In matters of health, variation in body temperature, pulse beat and blood

*Generally from the mean, infrequently from other measures of central tendency.

**The question of the *direction* of the variation will be discussed later in the text.

pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programmes. Thus measurement of variation is basic to the control of cause of variation. In engineering problems, measures of variation are often specially important. In social sciences, a special problem requiring the measurement of variability is the measurement of "inequality" of the distribution of income and wealth, etc.

(*iii*) Measures of variation enable comparison to be made of two or more series with regard to their variability. The study of variation may also be looked upon as a means of determining uniformity or consistency. A high degree of variation would mean little uniformity or consistency whereas a low degree of variation would mean greater uniformity or consistency.

(*iv*) Many powerful analytical tools in statistics such as correlation analysis, the testing of hypothesis, the analysis of fluctuations, techniques of production control, cost control, etc., are based on measures of variation of one kind or another.

Properties of a Good Measure of Variation

A good measure of variation should possess, as far as possible, the following properties :*

(*i*) It should be simple to understand.

(*ii*) It should be easy to compute.

(*iii*) It should be rigidly defined.

(*iv*) It should be based on each and every observation of the distribution.

(*v*) It should be amenable to further algebraic treatment.

(*vi*) It should have sampling stability.

(*vii*) It should not be unduly affected by extreme observations.

Methods of Studying Variation

The following are the important methods of studying variation :

I. The Range,

II. The Interquartile Range or Quartile Deviation,

III. The Average Deviation,

IV. The Standard Deviation, and

V. The Lorenz Curve.

Of these, the first four are mathematical methods and the last is a graphical one.

Absolute and Relative Measures of Variation

Measures of variation may be either absolute or relative. Absolute measures of variation are expressed in the same statistical unit in which the original data are given such as rupees, kilograms, tonnes, etc. These values may be used to compare the variation in two or more than two distributions provided the variables are expressed in the same units and have almost the same average value. In case the two sets of data are expressed in different units, such as quintals of sugar *versus* tonnes of sugarcane, or if the average value is very much different, such as manager's salary *versus* worker's salary, the absolute measures of variation are not comparable. In such cases measures of relative variation should be used.

*These properties are the same as those of a good mesure of central tendency. For details refer to the previous chapter.

A measure of relative variation is the ratio of a measure of absolute variation to an average. It is sometimes called a coefficient of variation, because "coefficient" means a pure number that is independent of the unit of measurement. It should be remembered that while computing the relative variation the average used as base should be the same one from which the absolute deviations were measured. This means that the arithmetic mean should be used with the standard deviation and either the arithmetic mean or median with the average deviation.

I. RANGE

Range is the simplest method of studying variation. It is defined as the difference between the value of the smallest observation and the value of the largest observation included in the distribution. Symbolically,

$$\text{Range} = L - S$$

L = Largest value, and

S = Smallest value

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula :

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

Illustration 1. The following are the prices of shares of a company from Monday to Saturday :

Day	*Price (Rs.)*	*Day*	*Price (Rs.)*
Monday	200	Thursday	160
Tuesday	210	Friday	220
Wednesday	208	Saturday	250

Calculate range and coefficient of range.

Solution.

$$\text{Range} = L - S$$

$L = 250$ and $S = 160$

$$\text{Range} = 250 - 160 = \text{Rs. } 90$$

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{250 - 160}{250 + 160} = \frac{90}{410} = 0.219$$

In a frequency distribution, range is calculated by taking the difference between the lower limit of the lowest class and the upper limit of the highest class.

Illustration 2. Calculate coefficient of range from the following data :

Profits (Rs. lakhs)	*No. of Cos.*	*Profits (Rs. lakhs)*	*No. of Cos.*
10–20	8	40–50	8
20–30	10	50–60	4
30–40	12		

Solution. Coefficient of Range $= \frac{L - S}{L + S} = \frac{60 - 10}{60 + 10} = \frac{50}{70} = 0.714.$

Merits and Limitations of Range

Merits. Among all the methods of studying variation, range is the simplest to understand and the easiest to compute. It takes minimum time to calculate the value of range. Hence, if one is interested in getting a quick rather than a very accurate picture of variability, one may compute range.

Limitations. (*i*) Range is not based on each and every observation of the distribution. (*ii*) It is subject to fluctuations of considerable magnitude from sample to sample. (*iii*) Range cannot be computed in case of open-end distributions. (*iv*) Range cannot tell us anything about the character of the distribution within two extreme observations. For example, observe the following three series:

Series A :	6,	46,	46,	46,	46,	46,	46,	46
Series B :	6,	6,	6,	6,	46,	46,	46,	46
Series C :	6,	10,	15,	25,	30,	32,	40,	46

In all the three series range is the same (*i.e.*, 46 – 6 = 40), but it does not mean that the distributions are alike. The range takes no account on the form of the distribution within the range. Range is, therefore, most unreliable as a guide to the variation of the values within a distribution.

Uses of Range

Despite serious limitations range is useful in the following cases :

(*i*) *Quality control.* The object of quality control is to keep a check on the quality of the product without 100% inspection. When statistical methods of quality control are used, control charts are prepared and in preparing these charts *range* plays a very important role. The idea basically is that if the range—the differences between the largest and smallest mass produced items—increases beyond a certain point, the production machinery should be examined to find out why the items of production have not followed their usual more consistent pattern.

(*ii*) *Fluctuation in the share prices.* Range is useful in studying the variations in the prices of stocks and shares and other commodities etc. They are very sensitive to price changes from one period to another.For example, by computing *range* we can get an idea about the range of variation of, say, gold prices. If the minimum price for 10 gm. during 2009–10 was Rs. 14,500 and the maximum price Rs. 16,700 this at once tells us about the range of variation, *i.e.*, 16,700 – 14,500 = 2,200.

(*iii*) *Weather forecasts.* The meteorological department does make use of the *range* in determining the difference between the minimum temperature and maximum temperature. This information is of great concern to the general public because they know as to within what limits the temperature is likely to vary on a particular day.

II. THE INTERQUARTILE RANGE OR QUARTILE DEVIATION

The range as a measure of variation has certain limitations. It is based on two extreme observations and it falls to take account of the scatter within the range. From this there is reason to believe that if the variation of the extreme observations is discarded the limited range thus, established might be more instructive. For this purpose there has been developed a measure called the *interquartile range*, the range which includes the middle 50 per cent of the observations. That is, one quartile of the observations at the lower end and another quartile of the observations at the upper end of the distribution are excluded in computing the inter-quartile range. In other words, interquartile range represents the difference between the third quartile and the first quartile. Symbolically,

$$\text{Interquartile range} = Q_3 - Q_1$$

Very often the interquartile range is reduced to the form of the semi-interquartile range or quartile deviation by dividing it by 2. Symbolically,

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

where $Q.D.$ = Quartile deviation.

Quartile deviation gives the average amount by which the two quartiles differ from the median. In asymmetrical distribution, the two quartiles (Q_1 and Q_3) are equidistant from the median, *i.e.*, Med.–Q_1 = Q_3–Med. and as such the difference can be taken as a measure of variation. The median ±$Q.D.$ covers exactly 50 per cent of the observations.

In reality, however, one seldom finds a series in business and economic data that is perfectly symmetrical. Nearly all distributions of social series are not equidistant from the median. As a result an asymmetrical distribution includes only approximately 50 per cent of the observations.

When quartile deviation is very small it describes high uniformity or small variation of the central 50% observations, and a high quartile deviation means that the variation among the central observations is large.

Quartile deviation is an absolute measure of variation. The relative measure corresponding to this measure, called the coefficient of quartile deviation, is calculated as follows :

$$\text{Coefficient of } Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Coefficient of quartile deviation can be used to compare the degree of variation in different distributions.

Computation of Quartile Deviation

The process of computing quartile deviation is very simple since we have just to compute the values of the upper and lower quartiles. The following illustration would clarify the procedure.

Illustration 3. You are given the frequency distribution of 292 workers of a factory according to their average weekly wages. Calculate quartile deviation and its coefficient from the following data :

Weekly wages (Rs.)	*No. of workers*	*Weekly wages (Rs.)*	*No. of workers*
Below 1350	8	1450–1470	22
1350–1370	16	1470–1490	15
1370–1390	39	1490–1510	15
1390–1410	58	1510–1530	9
1410–1430	60	1530 & above	10
1430–1450	40		

Solution. CALCULATION OF QUARTILE DEVIATION

Weekly wages (Rs.)	*No. of workers f*	*c.f.*
Below 1350	8	8
1350–1370	16	24
1370–1390	39	63
1390–1410	58	121
1410–1430	60	181
1430–1450	40	221
1450–1470	22	243
1470–1490	15	258
1490–1510	15	273
1510–1530	9	282
1530 & above	10	292
	$N = 292$	

$$\text{Med.} = \text{Size of } \frac{N}{2} \text{ th observation} = \frac{292}{2} = 146\text{th observation}$$

Median lies in the class 1410 – 1430.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i = 1410 + \frac{146 - 121}{60} \times 20 = 1410 + 8.333 = 1418.333$$

Q_1 = Size of $\frac{N}{4}$ th observation = $\frac{292}{4}$ = 73rd observation

Q_1 lies in the class 1390 –1410.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 1390 + \frac{73 - 63}{58} \times 20 = 1390 + 3.448 = 1393.448$$

Q_3 = Size of $\frac{3N}{4}$ th observation = $\frac{3 \times 292}{4}$ = 219 observation. Q_3 lies in the class 1430 – 1450.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i = 1430 + \frac{219 - 181}{40} \times 20 = 1430 + 19 = 1449$$

$$\text{Coeff. of } Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{1449 - 1393.448}{1449 + 1393.448} = \frac{55.552}{2842.448} = 0.020.$$

Merits and Limitations of Quartile Deviation

Merits. (*i*) In certain respects it is superior to range as a measure of variation. (*ii*) It has a special utility in measuring variation in case of open-end distributions or one in which the data may be ranked but measured quantitatively. (*iii*) It is also useful in erratic or highly skewed distributions, where the other measures of variation would be warped by extreme value. The quartile deviation is not affected by the presence of extreme values.*

Limitations. (*i*) Quartile deviation ignores 50% items, *i.e.*, the first 25% and the last 25%. As the value of quartile deviation does not depend upon every observation it cannot be regarded as a good method of measuring variation. (*ii*) It is not capable of mathematical manipulation. (*iii*) Its value is very much affected by sampling fluctuations. (*iv*) It is in fact not a measure of variation as it really does not show the scatter around an average but rather a distance on a scale, *i.e.*, quartile deviation is not itself measured from an average, but it is a positional average.

Consequently, some statisticians speak of quartile deviation as measure of *partition* rather than as a measure of variation. If we really desire to measure variation in the sense of showing the scatter around an average, we must include the deviation of each and every observation from an average in the measurement.

Because of the above limitations, quartile deviation is rarely used in practice.

III. THE AVERAGE DEVIATION

The two methods of variation discussed above, namely, range and quartile deviation, are not measures of variation in the strict sense of the term because they do not show the scatterness around an average. However, to study the formation of a distribution we should take the deviations from an average. The two other measures, namely, the average deviation and the standard deviation, help us in achieving this goal.

Average deviation is obtained by calculating the absolute deviations of each observation from median (or mean), and then averaging these deviations by taking their arithmetic mean. The formula for average deviation may be written as :

$$\text{A.D.}_{(\text{Med.})} = \frac{\Sigma |X - \text{Med.}|}{N}$$

In case deviations are taken from mean the formula shall be written as :

$$\text{A.D.}_{(\bar{X})} = \frac{\Sigma |X - \bar{X}|}{N}$$

*The range and *Q.D.* are positional measures of variation as they are based on the position of certain items in a distribution.

The reason for taking absolute deviations, that is, deviations in which signs are ignored, is that it is the amount of the differences of observations from median rather than the direction of the differences which is of main interest.

While calculating average deviation, deviation of observations can be taken from any average. However, theoretically speaking, there is an advantage in taking the deviations from median because the sum of the deviations of observations from median is minimum when signs are ignored. In actual practice the arithmetic mean is more popularly used in calculating the value of mean deviation because of its wide usage as a measure of central tendency. In any case, the average used must be clearly stated in a given problem so that any possible confusion in meaning is avoided.

Computation of Average Deviation—Ungrouped Data

The formula for computing average deviation is

$$\text{A.D.}_{(\text{Med.})} = \frac{\Sigma |X - \text{Med.}|}{N}$$

If the distribution is symmetrical the average (mean or median) $\pm$ average deviation is the range that will include 57.5 per cent of the observations in the series. If it is moderately skewed, then we may expect approximately 57.5 per cent of the observations to fall within this range. Hence if average deviation is small, the distribution is highly compact or uniform, since more than half of the cases are concentrated within a small range around the mean.

The relative measure corresponding to the average deviation, called the coefficient of average deviation, is obtained, by dividing average deviation by the particular average used in computing average deviation. Thus, if average deviation has been computed from median, the coefficient of average deviation shall be obtained by dividing average deviation by the median.

$$\text{Coefficient of A.D.}_{(\text{Med.})} = \frac{\text{A.D.}}{\text{Median}}$$

If mean has been used while calculating the value of average deviation, in such a case coefficient of average deviation shall be obtained by dividing average deviation by the mean.

Illustration 4. Calculate the average deviation and coefficient of average deviation of the two income groups of five and seven workers working in two different branches of a firm :

Branch I *Income (Rs.)*	*Branch II* *Income (Rs.)*
4,000	3,000
4,200	4,000
4,400	4,200
4,600	4,400
4,800	4,600
	4,800
	5,800

Solution. CALCULATION OF AVERAGE DEVIATION

Branch I		*Branch II*	
Income (Rs.)	$\lvert X - Med. \rvert$ *Med. = 4,400*	*Income (Rs.)*	$\lvert X - Med. \rvert$ *Med. = 4,400*
4,000	400	3,000	1,400
4,200	200	4,000	400
4,400	0	4,200	200
4,600	200	4,400	0
4,800	400	4,600	200
		4,800	400
		5,800	1,400
$N = 5$	$\Sigma \lvert X - Med. \rvert = 1,200$	$N = 7$	$\Sigma \lvert X - Med. \rvert = 4000$

Branch I : $$\text{A.D.} = \frac{\Sigma|X - \text{Med.}|}{N} = \frac{1200}{5} = 240.$$

$$\text{Coeff. of A.D.} = \frac{\text{A.D.}}{\text{Median}} = \frac{240}{4,400} = 0.054$$

Branch II : $$\text{A.D.} = \frac{\Sigma|X - \text{Med.}|}{N} = \frac{4,000}{7} = 571.43$$

$$\text{Coeff. of A.D.} = \frac{571.43}{4,400} = 0.13.$$

Calculation of Average Deviation—Grouped Data

In case of grouped data, the formula for calculating average deviation is :

$$\text{A.D.}_{(\text{Med.})} = \frac{\Sigma f\,|X - \text{Med.}|}{N}$$

Illustration 5. Calculate average deviation from mean from the following data :

Sales (in thousand Rs.)	*No. of days*	*Sales (in thousand Rs.)*	*No. of days*
10–20	3	40–50	3
20–30	6	50–60	2
30–40	11		

Solution. CALCULATION OF AVERAGE DEVIATION

Sales (in thousand Rs.)	*m.p. X*	*f*	*(X – 35)/10*	*fd*	$\lvert X - \bar{X} \rvert$	$f\lvert X - \bar{X} \rvert$
10–20	15	3	–2	–6	18	54
20–30	25	6	–1	–6	8	48
30–40	35	11	0	0	2	22
40–50	45	3	+1	+3	12	36
50–60	55	2	+2	+4	22	44
		$N = 25$		$\Sigma fd = -5$		$\Sigma f\lvert X - \bar{X} \rvert = 204$

$$\text{A.D.} = \frac{\Sigma f\,|X - \bar{X}|}{N}$$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 35 - \frac{5}{25} \times 10 = 35 - 2 = 33$$

$$\text{A.D.} = \frac{204}{25} = 8.16$$

Thus the average sales are Rs. 33 thousand per day and the average deviation of sales is Rs. 8.16 thousand.

Merits and Limitations of Average Deviation

Merits. (*i*) The outstanding advantage of the average deviation is its relative simplicity. It is simple to understand and easy to compute. Any one familiar with the concept of the average can readily appreciate the meaning of the average deviation. If a situation requires a measure of variation that will be presented to the general public or any group not thoroughly grounded in statistics, the average deviation is very useful. (*ii*) It is based on each and every observation of the data. Consequently change in the value of any observation would change the value of average deviation. (*iii*) Average deviation is less affected by the values of extreme observation. (*iv*) Since deviations are taken from a central value, comparison about formation of different distributions can easily be made.

Limitations. (*i*) The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items. If the signs of the deviations are not ignored, the net sum of the deviations will be zero if the reference point is the mean, or approximately zero if the reference point is

median. (*ii*) This method may not give us very accurate results. The reason is that average deviation gives us best results when deviations are taken from median. But median is not a satisfactory measure when the degree of variability in a series is very high. And if we compute average deviation from mean that is also not desirable because the sum of the deviations from mean (ignoring signs) is greater than the sum of the deviations from median (ignoring signs). If average deviation is computed from mode that also does not solve the problem because the value of mode cannot always be determined. (*iii*) It is not capable of further algebraic treatment. (*iv*) It is rarely used in sociological and business studies.

Because of these limitations its use is limited and it is overshadowed as a measure of variation by the superior standard deviation.

Usefulness of the Average Deviation. The serious drawbacks of the mean deviation should not blind us to its practical utility. Because of its simplicity in meaning and computation, it is especially effective in reports presented to the general public or to groups not familiar with statistical methods. This measure is useful for small samples with no elaborate analysis required. Incidentally it may be mentioned that the National Bureau of Economic Research has found in its work on forecasting business cycles, that the average deviation in the most practical measure of variation to use for this purpose.

IV. THE STANDARD DEVIATION

The standard deviation concept was introduced by Karl Pearson in 1893. It is by far the most important and widely used measure of studying variation. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of variation. It is a measure of how much "spread" or "variability" is present in the sample. If all the numbers in the sample are very close to each other, the standard deviation is close to zero.If the numbers are well dispersed, the standard deviation will tend to be large. Standard deviation is also known as *root mean square deviation* for the reason that it is the square root of the means of square deviations from the arithmetic mean. Standard deviation is denoted by the small Greek letter σ (read as sigma) and is defined as :

$$\sigma = \sqrt{\frac{\Sigma (X - \bar{X})^2}{N}}$$

If we square standard deviation, we get what is called Variance.

Hence Variance $= \sigma^2$ or $\sigma = \sqrt{\text{Variance}}$

The standard deviation measures the absolute variation of a distribution ; the greater the amount of variation, the greater the standard deviation, for the greater will be the magnitude of the deviation of the values from their mean. A small standard deviation means a high degree of uniformity of the observations as well as homogeneity of a series, a large standard deviation means just the opposite. Thus if we have two or more comparable series with identical or nearly identical means, it is the distribution with the smallest standard deviation that has the most representative mean. Hence standard deviation is extremely useful in judging the representativeness of the mean.

Calculation of Standard Deviation – Ungrouped Data

Standard deviation may be computed by applying any of the following two methods :

1. By taking deviations from the actual mean ; and
2. By taking deviations from an assumed mean.

1. Deviations taken from Actual Mean. When deviations are taken from the actual mean, the following formula is applied :

$$\sigma = \sqrt{\frac{\Sigma (X - \bar{X})^2}{N}} \quad ...(i)$$

If we calculate standard deviation without taking deviations, the above formula (*i*) after simplification (opening the brackets) can be used and is given by

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} \quad \text{or} \quad \sigma = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2}$$

2. Deviations taken from Assumed Mean. When the actual mean is in fractions, say 87.297, it would be too cumbersome to take deviations from it and then find squares of these deviations. In such a case either the mean may be approximated or else the deviations be taken from an assumed mean and the necessary adjustment be made in the value of standard deviation. The former method of approximation is less accurate and therefore, invariably in such a case deviations are taken from assumed mean.

When deviations are taken from assumed mean the following formula* is applied :

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2} \qquad ...(ii)$$

where $d = (X - A)$

Illustration 6. Find the standard deviation from the weekly wages of ten workers working in a factory :

Workers	*Weekly wages (Rs.)*	*Workers*	*Weekly wages (Rs.)*
A	1320	*F*	1340
B	1310	*G*	1325
C	1315	*H*	1321
D	1322	*I*	1320
E	1326	*J*	1331

Solution. CALCULATIONS OF STANDARD DEVIATION

Workers	*Weekly wages (Rs.)*	$(X - \bar{X})$	$(X - \bar{X})^2$
A	1320	–3	9
B	1310	–13	169
C	1315	–8	64
D	1322	–1	1
E	1326	+3	9

Contd.

*This formula has been derived from the original formula (*i*).

We know $\sigma = \sqrt{\frac{\Sigma (X - \bar{X})^2}{N}}$

Since $d = X - A, \therefore X = A + d$ and $\bar{X} = A + \bar{d}$

Subtracting $\bar{X}$ from X, we get

$$(X - \bar{X}) = (d - \bar{d})$$

Substituting the value of $(X - \bar{X})$ in (*i*), we have

$$\sigma = \sqrt{\frac{\Sigma (d - \bar{d})^2}{N}} = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

F	1340	+17	289
G	1325	+2	4
H	1321	–2	4
I	1320	–3	9
J	1331	+8	64
$N = 10$	$\Sigma X = 13230$	$\Sigma (X - \bar{X}) = 0$	$\Sigma (X - \bar{X})^2 = 622$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{13230}{10} = \text{Rs. } 1323$$

$$\sigma = \sqrt{\frac{\Sigma (X - \bar{X})^2}{N}} = \sqrt{\frac{622}{10}} = 7.89$$

If, in the above question, deviations are taken from 1320 instead of the actual mean 1323, the assumed mean method will be applied and the calculations would be as follows :

CALCULATION OF STANDARD DEVIATION (ASSUMED MEAN METHOD)

Workers	*Weekly wages (Rs.)* X	$(X - A)$ $A = 1320$ d	d^2
A	1320	0	0
B	1310	–10	100
C	1315	–5	25
D	1322	+2	4
E	1326	+6	36
F	1340	+20	400
G	1325	+5	25
H	1321	+1	1
I	1320	0	0
J	1331	+11	121
$N = 10$		$\Sigma d = 30$	$\Sigma d^2 = 712$

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2} = \sqrt{\frac{712}{10} - \left(\frac{30}{10}\right)^2} = \sqrt{71.2 - 9} = \sqrt{62.2} = 7.89$$

Thus the answer remains the same by both the methods. It should be noted that when actual mean is not a whole number, assumed mean method should be preferred because it simplifies calculations.

Calculation of Standard Deviation—Grouped Data

In grouped frequency distribution, standard deviation can be calculated by applying any of the following two methods :

1. By taking deviations from actual mean; and
2. By taking deviations from assumed mean.

1. Deviations taken from Actual Mean. When deviations are taken from actual mean, the following formula* is used :

$$\sigma = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}}$$

If we calculate standard deviation without taking deviations, then this formula after simplification (opening the brackets) can be used and is given by

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \left(\frac{\Sigma fX}{N}\right)^2} \quad \text{or} \quad \sigma = \sqrt{\frac{\Sigma fX^2}{N} - (\bar{X})^2}$$

2. Deviations taken from Assumed Mean. When deviations are taken from the assumed mean, the following formula* is applied :

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

Illustration 7. An analysis of production rejects resulted in the following figures :

No. of rejects per operator	*No. of operators*	*No. of rejects per operator*	*No. of operators*
21–25	5	41–45	15
26–30	15	46–50	12
31–35	28	51–55	3
36–40	42		

Calculate mean and standard deviation.

Solution. Converting the discrete data into the continuous data, we get the following table :

CALCULATION OF MEAN AND STANDARD DEVIATION

No. of rejects per operator	*m.p. X*	*No. of operators f*	*(X – 38)/5 d*	*fd*	*fd²*
20.5–25.5	23	5	–3	–15	45
25.5–30.5	28	15	–2	–30	60
30.5–35.5	33	28	–1	–28	28
35.5–40.5	38	42	0	0	0
40.5–45.5	43	15	+1	+15	15
45.5–50.5	48	12	+2	+24	48
50.5–55.5	53	3	+3	+9	27
		$N = 120$		$\Sigma fd = -25$	$\Sigma fd^2 = 223$

* $$\sigma = \frac{\Sigma f(X-\bar{X})^2}{N} \qquad ...(i)$$

Let $d = \frac{X-A}{i}$; then $X = A + id$ and $\bar{X} = A + i\bar{d}$

Substituting the value of $(X - \bar{X})$ in (i), we have

$$\sigma = \sqrt{\frac{\Sigma f[i(d-\bar{d})]^2}{N}} = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$\text{Mean : } \bar{X} = A + \frac{\Sigma fd}{N} \times i = 38 - \frac{25}{120} \times 5 = 38 - 1.04 = 36.96$$

$$\text{Standard deviation : } \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{223}{120} - \left(\frac{-25}{120}\right)^2} \times 5$$

$$= \sqrt{1.858 - .043} \times 5 = 1.347 \times 5 = 6.375.$$

Illustration 8. An association doing charity work decided to give old age pensions to people over sixty years of age.

The scales of pensions were fixed as follows :

Age group 60 to 65—Rs. 2500 per month.

Age group 65 to 70—Rs. 3000 per month.

Age group 70 to 75—Rs. 3500 per month.

Age group 75 to 80—Rs. 4000 per month.

Age group 80 to 85—Rs. 4500 per month.

The age of 25 persons who secured the pension benefits are given below :

75	62	84	72	83	72	81	64	71	63	61	60	61
67	74	64	79	73	75	76	69	78	66	67	68	

Calculate the monthly average pension payable and the standard deviation.

Solution. CLASSIFYING THE ABOVE DATA

Age Group	*Tally*	*Frequency*
60–65	𝍸 ||	7
65–70	𝍸	5
70–75	𝍸 |	6
75–80	||||	4
80–85	|||	3
		$N = 25$

CALCULATIONS OF MONTHLY AVERAGE PENSION PAYABLE AND THE STANDARD DEVIATION

Pension (Rs.) X	$(X - 3500)/500$ d	f	fd	fd^2
2500	−2	7	−14	28
3000	−1	5	−5	5
3500	0	6	0	0
4000	+1	4	+4	4
4500	+2	3	+6	12
		$N = 25$	$\Sigma fd = -9$	$\Sigma fd^2 = 49$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 3500 - \frac{9}{25} \times 500 = 3500 - 180 = \text{Rs. } 3320$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{49}{25} - \left(\frac{-9}{25}\right)^2} \times 500 = 1.353 \times 500 = \text{Rs. } 676.5$$

Thus the monthly average pension is Rs. 3320 and standard deviation Rs. 676.5.

Mathematical Properties of Standard Deviation

Standard deviation has some very important mathematical properties which considerably enhance its utility in statistical work.

1. *Combined Standard Deviation.* Just as it is possible to compute combined mean of two or more than two groups, similarly we can also compute combined standard deviation of two or more groups. Combined standard deviation of two groups is denoted by σ_{12} and is computed as follows :

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

where σ_{12} = combined standard deviation,

σ_1 = standard deviation of first group,

σ_2 = standard deviation of second group, and

$d_1 = |\bar{X}_1 - \bar{X}_{12}|$; $d_2 = |\bar{X}_2 - \bar{X}_{12}|$.

The above formula can be extended to find out the standard deviation of three or more groups. For example, combined standard deviation of three groups would be :

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

where $d_1 = |\bar{X}_1 - \bar{X}_{123}|$; $d_2 = |\bar{X}_2 - \bar{X}_{123}|$; $d_3 = |\bar{X}_3 - \bar{X}_{123}|$.

Illustration 9. The number of workers employed, the mean wage (in Rs.) per week and the standard deviation (in Rs.) in each branch of a company are given below. Calculate mean wages and standard deviation of all the workers taken together for company.

Branch	*No. of workers employed*	*Weekly mean wage (in Rs.)*	*Standard deviation (in Rs.)*
A	50	1413	60
B	60	1420	70
C	90	1415	80

Solution :

$$\bar{X}_{123} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3}$$

$$= \frac{(50 \times 1413) + (60 \times 1420) + (90 \times 1415)}{50 + 60 + 90}$$

$$= \frac{70{,}650 + 85{,}200 + 1{,}27{,}350}{200} = \frac{2{,}83{,}200}{200} = \text{Rs. } 1{,}416$$

Combined standard deviation of three branches

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

$$d_1 = \left|\bar{X}_1 - \bar{X}_{123}\right| = |1413 - 1416| = 3$$

$$d_2 = \left|\bar{X}_2 - \bar{X}_{123}\right| = |1420 - 1416| = 4$$

$$d_3 = \left|\bar{X}_3 - \bar{X}_{123}\right| = |1415 - 1416| = 1$$

$$\sigma_{123} = \sqrt{\frac{50(60)^2 + 60(70)^2 + 90(80)^2 + 50(3)^2 + 60(4)^2 + 90(1)^2}{50 + 60 + 90}}$$

$$= \sqrt{\frac{1,80,000 + 2,94,000 + 5,76,000 + 450 + 960 + 90}{200}}$$

$$= \sqrt{\frac{1051500}{200}} = \text{Rs. } 72.51.$$

2. *Standard deviation of natural numbers.* The standard deviation of the first n natural numbers* can be obtained by the following formula :

$$\sigma = \sqrt{\frac{1}{12}(N^2 - 1)}$$

Thus the standard deviation of natural numbers 1 to 10 will be

$$\sigma = \sqrt{\frac{1}{12}(10^2 - 1)} = \sqrt{\frac{1}{12} \times 99} = \sqrt{8.25} = 2.87.$$

3. The sum of the squares of the deviations of all the observations from their arithmetic mean is minimum. In other words, the sum of the squares of the deviations of observations from a value other than the arithmetic mean would always be greater. This is the reason why standard deviation is always computed from the arithmetic mean.

4. Standard deviation is independent of change of origin but not scale.

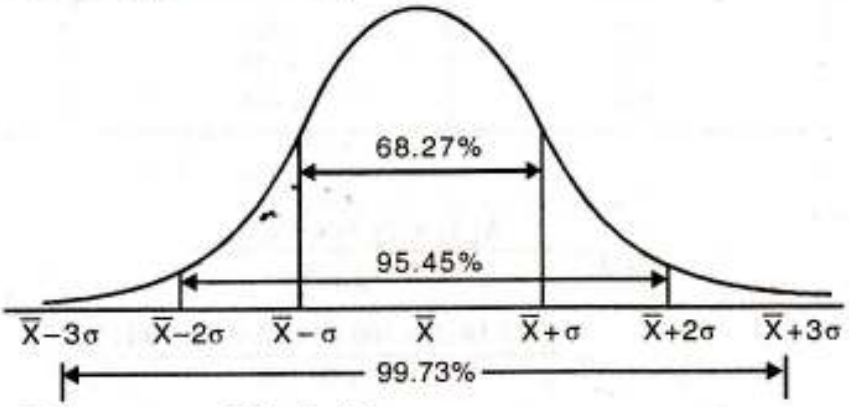
5. For a symmetrical distribution, the following area relationships hold good.

Mean $\pm$ 1 σ covers 68.27% observations.

Mean $\pm$ 2 σ covers 95.45% observations.

Mean $\pm$ 3 σ covers 99.73% observations.

This can be illustrated by the following diagram :



Relation between Measures of Variation

In a bell-shaped symmetrical distribution, there is a fixed relationship between the three most commonly used measures of variation. The quartile deviation is smallest, the mean deviation next and the standard deviation is largest in the following proportions :

$$Q.D. = \frac{2}{3}\sigma \text{ ; and } A.D. = \frac{4}{5}\sigma$$

These relationships can be easily memorized because of the sequence 2, 3, 4, 5. The same proportions tend to hold true for many distributions that are quite symmetrical. They are useful in estimating one measure of variation when another is known, or in checking roughly the accuracy of a calculated value. If the computed σ differs very widely from its value estimated from $Q.D.$ or $A.D.$ either an error has been made or the distribution differs considerably from symmetry.

*By natural number we mean only positive integers, *i.e.*, 1, 2, 3, 4, 5 ...

Another comparison may be made of the proportion of observations that are 'typically included within the range of one *Q.D.*, *A.D.* or *S.D.* measured both above and below the mean. In a normal distribution :

$\overline{X} \pm Q.D.$ includes 50 per cent of the observations.

$\overline{X} \pm A.D.$ includes 57.51 per cent of the observations.

$\overline{X} \pm \sigma$ includes 68.27 per cent or about two-thirds of the observations.

Illustration 10. The breaking strength of 80 'test pieces' of a certain alloy is given in the following table, the unit being given to the nearest thousand pounds per square inch.

Breaking strength	*No. of pieces*
44–46	3
46–48	24
48–50	27
50–52	21
52–54	5

Calculate the average breaking strength of the alloy and the standard deviation. Calculate the percentage of observations lying between mean $\pm 2\sigma$. *(MBA, Kurukshetra Univ; MBA, Madras Univ., 2006)*

Solution : CALCULATION OF MEAN AND STANDARD DEVIATION

Breaking strength	*m.p.* X	f	$(X-49)/2$ d	fd	fd^2
44–46	45	3	−2	−6	12
46–48	47	24	−1	−24	24
48–50	49	27	0	0	0
50–52	51	21	+1	+21	21
52–54	53	5	+2	+10	20
		$N = 80$		$\Sigma fd = 1$	$\Sigma fd^2 = 77$

$$\overline{X} = A + \frac{\Sigma fd}{N} \times i = 49 + \frac{1}{81} \times 2 = 49.025$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{77}{80} - \left(\frac{1}{80}\right)^2} \times 2$$

$$= \sqrt{0.962 - 0.0001} \times 2 = 0.98 \times 2 = 1.96$$

Hence $\overline{X} = 49.025, \sigma = 1.96$

The value of $\overline{X} \pm 2\sigma = 49.025 \pm 2\,(1.96) = 49.025 \pm 3.92$

or $= 45.105, 52.945 = 45$ and 53 approx.

We have to calculate the percentage of items lying between 45 and 53. For this we make an assumption that the number of pieces are equally distributed within each class. Since between 44 and 46 there are 3 frequencies at 45, these would be 1.5. Similarly at 53 the frequency would be 2.5. Thus the total frequency between 45 and 53 = (1.5 + 24 + 27 + 21 + 2.5) = 76. The percentage is $\frac{76}{80} \times 100 = 95$. Thus there are 95 per cent observations lying within the limits mean $\pm 2\sigma$.

*For details please refer to Chapter on 'Probability Distribution'.

Merits and Limitations of Standard Deviation

Merits. (*i*) The standard deviation is the best measure of variation because of its mathematical characteristics. It is based on every item of the distribution. Also it is amenable to algebraic treatment and is less affected by fluctuations of sampling than most other measures of variation.

(*ii*) It is possible to calculate the combined standard deviation of two or more groups. This is not possible with any other measure.

(*iii*) For comparing the variability of two or more distributions coefficient of variation is considered to be most appropriate and this measure is based on mean and standard deviation.

(*iv*) Standard deviation is most prominently used in further statistical work. For example, in comparing skewness, correlation, etc., use is made of standard deviation. It is a key-note in sampling and provides a unit of measurement for the normal distribution.

Limitations. (*i*) As compared to other measures it is difficult to compute. However, it does not reduce the importance of this measure because of the high degree of accuracy of result it gives.

(*ii*) It gives more weight to extreme values and less to those which are near the mean. It is because of the fact that the squares of the deviations which are big in size would be proportionately greater than the squares of those deviations which are comparatively small. The deviations 2 and 8 are in ratio of 1 : 4 but their squares, *i.e.*, 4 and 64, would be in the ratio 1 : 16.

Correcting Incorrect Value of Standard Deviation

Mistakes in calculations are always possible. Sometimes it so happens that while calculating mean and standard deviation we unconsciously copy out wrong observations. For example, an observation 21 may be poised as 12. Similarly, one observation 127 may be taken as only 27. In such a case if the entire calculations are done again, it would become too difficult a task. By adopting a very simple procedure we can correct the incorrect values of mean and standard deviation. For obtaining correct mean we find out correct ΣX by deducting from the original ΣX the wrong observations and adding to it the correct observations. Similarly, for calculating correct standard deviation we obtain the value of correct ΣX^2. The following illustration shall clarify the procedure :

Illustration 11. The mean and standard deviation of a set of 100 observations were worked out as 40 and 5 respectively by a computer which by mistake took the value 50 in place of 40 for one of the observations. Find the correct mean and variance.

(M. Com., Jammu Univ. ; MBA, Lucknow Univ., 2002)

Solution :

$$\bar{X} = \frac{\Sigma X}{N}, \quad N\bar{X} = \Sigma X, N = 100, \ \bar{X} = 40$$

$$\therefore \quad \Sigma X = 100 \times 40 = 4{,}000$$

But this is not the correct ΣX because one observation has been taken as 50 instead of 40.

$$\therefore \quad \text{Correct } \Sigma X = 4{,}000 - 50 + 40 = 3{,}990$$

$$\text{Correct Mean} = \frac{3{,}990}{100} = 39.9$$

$$\text{Variance} = \frac{\Sigma X^2}{N} - (\bar{X})^2$$

$$\text{Variance} = \sigma^2 = (5)^2 = 25, \ N = 100.$$

$$25 = \frac{\Sigma X^2}{100} - (40)^2 \quad \text{or} \quad 2{,}500 = \Sigma X^2 - 1{,}60{,}000$$

$$\Sigma X^2 = 1{,}60{,}000 + 2{,}500 = 1{,}62{,}500$$

$$\text{Correct } \Sigma X^2 = 1{,}62{,}500 - (50)^2 + (40)^2 = 1{,}62{,}500 - 2{,}500 + 1{,}600 = 1{,}61{,}600$$

$$\text{Correct variance} = \frac{\text{Correct } \Sigma X^2}{N} - (\textit{Correct } \bar{X})^2$$

$$= \frac{1{,}61{,}600}{100} - (39.9)^2 = 1{,}616 - 1{,}592.01 = 23.99$$

Thus correct mean = 39.9 and the correct variance = 23.99.

Coefficient of Variation

The standard deviation discussed so far is an absolute measure of variation. The corresponding relative measure is known as the *coefficient of variation.* This measure developed by Karl Pearson is the most commonly used measure of relative variation. It is used in such problems where we want to compare the variability of two or more than two series. That series (or group) for which the coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous. On the other hand, the series for which coefficient of variation is less is said to be less variable or more consistent, more uniform, more stable or more homogeneous. Coefficient of variation denoted by C.V. is obtained as follows :

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100$$

It may be pointed out that although any measure of variation can be used in conjunction with any average in computing relative variation, statisticians, in fact, almost always use the standard deviation as the measure of variation and the arithmetic mean as the average. Coefficient of variation is more useful when the two distributions are entirely different and the units of measurement are also different. When the relative variation is stated in terms of the arithmetic mean and the standard deviation, the resulting percentage is known as the coefficient of variation or coefficient of variability.

Illustration 12. Suppose that samples of polythene bags from two manufacturers *A* and *B* are tested by a prospective buyer for bursting pressure, with the following results :

Bursting pressure (*lbs*)	*Number of bags*	
	A	*B*
5.0–9.9	2	9
10.0–14.9	9	11
15.0–19.9	29	18
20.0–24.9	54	32
25.0–29.9	11	27
30.0–34.9	5	13
	110	110

Which set of bags has the highest average bursting pressure? Which has more uniform pressure? If prices are the same, which manufacturer's bags would be preferred by the buyer? Why?

Solution. For determining which set of bags has the highest average bursting pressure, calculate arithmetic mean and for finding out which has more uniform pressure compute coefficient of variations.

Manufacturer A :

CALCULATION OF MEAN AND STANDARD DEVIATION

Bursting pressure (*lbs*)	*m.p.* *X*	*f*	(*X*–17.45)/5 *d*	*fd*	fd^2
4.95–9.95	7.45	2	–2	–4	8
9.95–14.95	12.45	9	–1	–9	9
14.95–19.95	**17.45**	29	0	0	0
19.95–24.95	22.45	54	+1	+54	54
24.95–29.95	27.45	11	+2	+22	44
29.95–34.95	32.45	5	+3	+15	45
		$N = 110$		$\Sigma fd = 78$	$\Sigma fd^2 = 160$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 17.45 + \frac{78}{110} \times 5 = 17.45 + 3.55 = 21$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{160}{110} - \left(\frac{78}{110}\right)^2} \times 5$$

$$= \sqrt{1.455 - 0.503} \times 5 = \sqrt{0.952} \times 5 = 0.976 \times 5 = 4.879$$

$$C.V. = \frac{\sigma}{\bar{X}} \times 100 = \frac{4.879}{21} \times 100 = 23.23\%.$$

Manufacturer B :

CALCULATION OF MEAN AND STANDARD DEVIATION

Bursting pressure (lbs.)	*m.p. X*	*f*	*(X–17.45)/5 d*	*fd*	fd^2
4.95–9.95	7.45	9	–2	–18	36
9.95–14.95	12.45	11	–1	–11	11
14.95–19.95	17.45	18	0	0	0
19.95–24.95	22.45	32	+1	+32	32
24.95–29.95	27.45	27	+2	+54	108
29.95–34.95	32.45	13	+3	+39	117
		$N = 110$		$\Sigma fd = 96$	$\Sigma fd^2 = 304$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 17.45 + \frac{96}{110} \times 5 = 21.81$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{304}{110} - \left(\frac{96}{110}\right)^2} \times 5$$

$$= \sqrt{2.764 - 0.762} \times 5 = 1.4149 \times 5 = 7.0745$$

$$C.V. = \frac{\sigma}{\bar{X}} \times 100 = \frac{7.0745}{21.81} \times 100 = 32.44\%$$

Since the average bursting pressure is higher for manufacturer *B*, hence the bags of manufacturer *B* have a higher bursting pressure. The bags of manufacturer *A* have more uniform pressure since the coefficient of variation is less for manufacturer *A*. If prices are the same, the bags of manufacturer *A* should be preferred by the buyer because they have more uniform pressure.

Illustration 13. In two factories *A* and *B* engaged in the same industry, the average monthly wages and standard deviations are as follows :

Factory	*Average Monthly Wages (Rs.)*	*S.D. of Wages (Rs.)*	*No. of Wage Earners*
A	4600	500	100
B	4900	400	80

(*i*) Which factory *A* or *B* pays larger amount as monthly wages?

(*ii*) Which factory shows greater variability in the distribution of wages?

(*iii*) What is the mean and standard deviation of all the workers in two factories taken together?

Solution. (*i*) For finding out which factory *A* or *B* pays larger amount as monthly wages; we have to compare the total wage bill.

Factory A : Total wage bill = 4600 × 100 = Rs. 4,60,000

Factory B : Total wage bill = 4900 × 80 = Rs. 3,92,000

Hence factory *A* pays larger amount as monthly wages.

(*ii*) For determining which factory shows greater variation in the distribution of wages, we have to compare coefficient of variation.

$$C.V. \text{ (factory } A) = \frac{\sigma}{\bar{X}} \times 100 = \frac{500}{4600} \times 100 = 10.87$$

$$C.V. \text{ (factory } B) = \frac{\sigma}{\bar{X}} \times 100 = \frac{400}{4900} \times 100 = 8.16$$

Since coefficient of variation is higher in factory *A*, hence factory *A* shows greater variability in the distribution of wages.

(*iii*) ***Combined mean*** :

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$= \frac{(100 \times 4600) + (80 \times 4900)}{100 + 80} = \frac{460000 + 392000}{180}$$

$$= \frac{852000}{180} = \text{Rs. } 4,733.33$$

Combined Standard Deviation : $\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$

$$N_1 = 100, \sigma_1 = 500, N_2 = 80, \sigma_2 = 400, d_1 = |\bar{X}_2 - \bar{X}_{12}|$$

$$|4600 - 4733.33| = 133.33, d_2 = |\bar{X}_2 - \bar{X}_{12}| = |4900 - 4733.33| = 166.67$$

$$\sigma_{12} = \sqrt{\frac{100\,(500)^2 + 80\,(400)^2 + 100\,(133.33)^2 + 80\,(166.67)^2}{100 + 80}}$$

$$= \sqrt{\frac{25000000 + 1200000 + 177768.89 + 222231.11}{180}}$$

$$= \sqrt{\frac{42192486.335}{180}} = 484.15.$$

Hence combined standard deviation is Rs. 484.15.

Illustration 14. The number of employees, daily wages per employee and the variance of the wages per employee for two factories are given below :

	Factory A	*Factory B*
Number of employees	50	100
Average daily wages per employee (Rs.)	120	85
Variance of the daily wages per employee (Rs.)	9	16

(*a*) In which factory is there greater variation in the distribution of daily wages per employee?

(*b*) Suppose in factory *B*, the daily wages of an employee were wrongly noted as Rs. 120 instead of Rs. 100. What would be the correct variance for factory *B*? ***(Diploma in Mgt., IGNOU, 2002; MBA, UP Tech, Univ, 2003)***

Solution : (*a*) *Variation in the Distribution of Wages*

Factory A

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

$$\sigma = \sqrt{9} = 3, \bar{X} = 120$$

$$\therefore \quad C.V. = \frac{3}{120} \times 100 = 2.5$$

Factory B

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

$$\sigma = \sqrt{16} = 4, \bar{X} = 85$$

$$\therefore C.V. = \frac{4}{85} \times 100 = 4.7$$

The coefficient of variation is greater for factory *B*, hence there is greater variation in the distribution of wages per employee in factory *B*.

(*b*) *Correct Variance* : For finding correct variance we have first to find the correct mean.

$$\bar{X} = \frac{\Sigma X}{N}, N\bar{X} = \Sigma X$$

$$\Sigma X = 100 \times 85 = 8,500$$

$$\text{Correct } \Sigma X = 8,500 - 120 + 100 = 8,480$$

$$\text{Correct mean} = \frac{8480}{100} = 84.8$$

$$\text{Variance or } \sigma^2 = \frac{\Sigma X^2}{N} - (\bar{X})^2$$

Substituting the values of σ^2, $\bar{X}$, etc.,

$$16 = \frac{\Sigma X^2}{100} - (85)^2 \text{ or } 1600 = \Sigma X^2 - 7,22,500$$

$$\Sigma X^2 = 7,24,100$$

But in this total 100 has been taken as 120.

$$\therefore \quad \text{Correct } \Sigma X^2 = 7,24,100 - (120)^2 + (100)^2$$

$$= 7,24,100 - 14,400 + 10,000 = 7,19,700$$

$$\text{Correct variance} = \frac{\text{Correct } \Sigma X^2}{N} - (\text{correct } \bar{X})^2$$

$$= \frac{7,19,700}{100} - (84.8)^2 = 7,197 - 7,191.04 = 5.96$$

Hence the correct variance for factory *B* is 5.96.

V. LORENZ CURVE

The Lorenz Curve, devised by Max O. Lorenz, a famous economic statistician, is a graphic method of studying variation. This curve was used by him for the first time to measure the distribution of wealth and income. Now the curve is also used to study the distribution of profits, wages, turnover, etc. However, still the most common use of this curve is in the study of the degree of inequality in the distribution of income and wealth between countries or between different periods of time. It is a cumulative percentage curve in which the percentage of items is combined with the percentage of other things as wealth, profits, turnover, etc.

While drawing the Lorenz Curve the following procedure is adopted :

(*i*) The size of items and frequencies are both cumulated and then percentages are obtained for these various cumulative values.

(*ii*) On the *X*-axis, start from 0 to 100 and take the per cent of variable.

(*iii*) On the *Y*-axis, start from 0 to 100 and take the per cent of variable.

(*iv*) Draw a diagonal line joining 0 with 100. This is known as line of equal distribution. Any point on this line shows the same per cent on *X* as on *Y*.

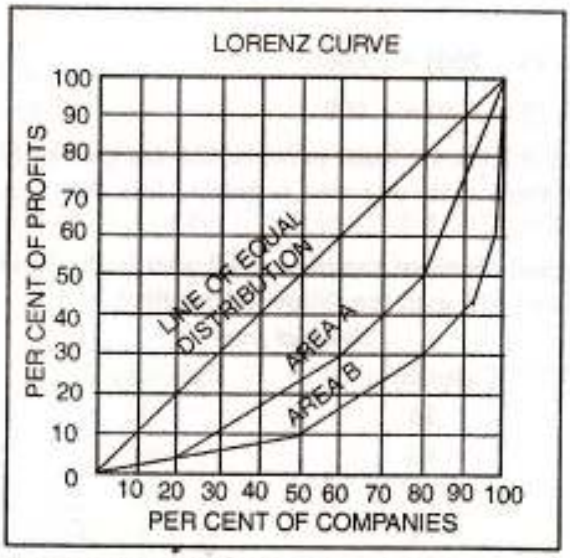
(*v*) Plot the various points corresponding to *X* and *Y* and join them. The distribution so obtained, unless it is exactly equal, will always curve below the diagonal line. If two curves of distribution are shown on the same Lorenz presentation, the curve that is farthest from the diagonal line represents the greater inequality. Clearly the line of actual distribution can never cross the line of equal distribution.

Illustration 15. In the following table is given the number of companies belonging to two areas *A* and *B* according to the amount of profits earned by them. Draw in the same diagram their Lorenz curves and interpret them.

Profits earned in	*No. of Companies*	
Rs. '000	*Area A*	*Area B*
6	6	2
25	11	38
60	13	52
84	14	28
105	15	38
150	17	26
170	10	12
400	14	4

Solution. CALCULATION FOR DRAWING THE LORENZ CURVE

	Profit		*Area A*			*Area B*		
Profits earned in Rs. '000	*Cumulative profits*	*Cumulative percentage*	*No. of Companies*	*Cumulative Number*	*Cumulative Percentage*	*No. of Companies*	*Cumulative Number*	*Cumulative Percentage*
6	6	0.6	6	6	6	2	2	1
25	31	3.1	11	17	17	38	40	20
60	91	9.1	13	30	30	52	92	46
84	175	17.5	14	44	44	28	120	60
105	280	28.0	15	59	59	38	158	79
150	430	43.0	17	76	76	26	184	92
170	600	60.0	10	86	86	12	196	98
400	1000	100.0	14	100	100	4	200	100



Since curve *B* is farther from the diagonal line, it represents greater inequality.

Which Measure of Variation to use ?

Unlike measures of central tendency, in case of measures of variation also the question arises which measure to use. The choice of a suitable measure depends on the following three factors :

1. *The type of data available.* If observations are few in numbers, or contain extreme values, avoid the standard deviation. If they are generally skewed, avoid the mean deviation as well. If they have gaps around the quartiles, the quartile deviation should be avoided. If there are open-end classes, the quartile measure of variation should be preferred.

2. *The purpose of investigation.* In an elementary treatment of statistical series in which a measure of variability is desired only for itself, any of the three measures, namely, range, quartile deviation and average deviation, would be acceptable. Probably the average deviation would be superior. However, in usual practice, the measure of variability is employed in further statistical analysis. For such a purpose, the standard deviation is by far the most popularly used. It is free from those defects with which other measures suffer. It lends itself to the analysis of variability in terms of normal curve of error. Practically, all advanced statistical methods deal with variability and centre around the standard deviation. Hence unless the circumstances warrant for the use of any other measure, we should make use of standard deviation for measuring variability.

MISCELLANEOUS ILLUSTRATIONS

Illustration 16. You are incharge of rationing in a State affected by food shortage. The following reports arrive from local investigators :

Daily caloric value of food available per adult during current period :

Area	*Mean*	*Standard Deviation*
A	2,500	400
B	2,000	200

The estimated requirement of an adult is taken at 2,800 calories daily and the absolute minimum is 1,350. Comment on the reported figures, and determine which area, in your opinion, need more urgent attention.

Solution. We know that $\bar{X} \pm 1\sigma$ covers 68.2% cases ; $\bar{X} \pm 2\sigma$ covers 95.45% cases and $\bar{X} \pm 3\sigma$ covers 93.73% cases. In the given problem, if we take into consideration 99.73%, *i.e.*, almost the whole of the population, the limits would be $\bar{X} \pm 3\sigma$.

For Area *A*, these limits are :

$$\bar{X} + 3\sigma = 2,500 + (3 \times 400) = 3,700$$

$$\bar{X} - 3\sigma = 2,500 - (3 \times 400) = 1,300$$

For Area *B*, these limits are :

$$\bar{X} + 3\sigma = 2,000 + (3 \times 200) = 2,600$$

$$\bar{X} - 3\sigma = 2,000 - (3 \times 200) = 1,400$$

It is clear from above that in Area *A* there are some persons who are getting 1300 calories, *i.e.*, below the minimum which is 1,350. But in case of area *B* there is no one who is getting less than the minimum. Hence area *A* needs more urgent attention.

Illustration 17. A purchasing agent obtained samples of 60 watt bulbs from two companies. He had the samples tested in his own laboratory for length of life with the following results :

Length of life (in hours)	*Samples from*	
	Company A	*Company B*
1,700 and under 1,900	10	3
1,900 " " 2,100	16	40
2,100 " " 2,300	20	12
2,300 " " 2,500	8	3
2,500 " " 2,700	6	2

(*a*) Which Company's bulbs do you think are better in terms of average life ?

(*b*) If prices of both types are the same, which company's bulbs would you buy and why ? *(MBA, DU, 2000)*

Solution. In order to answer these questions we have to calculate mean and coefficient of variation.

CALCULATION OF MEAN AND COEFFICIENT OF VARIATION

Length of life (in hours)	*m.p.* X	$(X-2200)/200$ d	*Samples from Co. A* f	fd	fd^2	*Samples from Co. B* f	fd	fd^2
1,700–1,900	1,800	–2	10	–20	40	3	–6	12
1,900–2,100	2,000	–1	16	–16	16	40	–40	40
2,100–2,300	**2,200**	0	20	0	0	12	0	0
2,300–2,500	2,400	+1	8	+8	8	3	+3	3
2,500–2,700	2,600	+2	6	+12	24	2	+4	8
			$N = 60$	$\Sigma fd = -16$	$\Sigma fd^2 = 88$	$N = 60$	$\Sigma fd = -39$	$\Sigma fd^2 = 63$

For Company A :

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 2,200 - \frac{16}{60} \times 200 = 2,146.67$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{88}{60} - \left(\frac{-16}{60}\right)^2} \times 200$$

$$= \sqrt{1.467 - 0.071} \times 200 = 1.182 \times 200 = 236.4$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{236.4}{2146.67} \times 100 = 11 \text{ per cent.}$$

For Company B :

$$\bar{X} = 2200 - \frac{39}{60} \times 200 = 2200 - 130 = 2070$$

$$\sigma = \sqrt{\frac{63}{60} - \left(\frac{-39}{60}\right)^2} \times 200 = \sqrt{1.05 - .42} \times 200 = .794 \times 200 = 158.8$$

$$\text{C.V.} = \frac{158.8}{2070} \times 100 = 7.67 \text{ per cent.}$$

(*a*) Since average length of life is greater in case of company *A*, hence bulbs of company *A* are better.

(*b*) Coefficient of variation is less for company *B*. Hence if prices are same, we will prefer to buy company *B*'s bulbs because their burning hours are more uniform.

Illustration 18. You are given the data pertaining to kilowatt hours of electricity consumed by 100 persons in Delhi.

Consumption (K. Watt hours)	*No. of users*
0 but less than 10	6
10 " " " 20	25
20" " " 30	36
30 " " " 40	20
40 " " " 50	13

Calculate (*i*) the standard deviation, and (*iii*) the range within which middle 50% of the consumers fall.

Solution. CALCULATION OF MEAN AND STANDARD DEVIATION

Consumption K. watt hours	*m.p. X*	*No. of Users f*	*(X–25)/10 d*	*fd*	fd^2	*c.f.*
0–10	5	6	–2	–12	24	6
10–20	15	25	–1	–25	25	31
20–30	**25**	36	0	0	0	67
30–40	35	20	+1	+20	20	87
40–50	45	13	+2	+26	52	100
		N=100		Σfd=9	Σfd^2=121	

(*i*)
$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 25 + \frac{9}{100} \times 10 = 25.9 \text{ k. watt hours}$$

(*ii*)
$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{121}{100} - \left(\frac{9}{100}\right)^2} \times 10$$

$$= \sqrt{1.21 - .008} \times 10 = 1.096 \times 10 = 10.96$$

(*iii*) For calculating the range, we have to find Q_1 and Q_3.

Q_1 = size of $\frac{N}{4}$ th observation = $\frac{100}{4}$ = 25th observation,

Q_1 lies in the class 10 – 20.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 10 + \frac{25-6}{25} \times 10 = 10 + 7.6 = 17.6$$

Q_3 = size of $\frac{3N}{4}$ th observation = $\frac{3 \times 100}{4}$ = 75th observation

Q_3 lies in the class 30–40.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i = 30 + \frac{75-67}{20} \times 10 = 30 + 4 = 34$$

Range within which the middle 50% of the consumers fall

$$= Q_3 - Q_1 = 34 - 17.6 = 16.4.$$

Illustration 19. The value of the arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from the use of working origin and scale are Rs. 107 and 13.1 respectively. Determine the actual classes.

d:	−3	−2	−1	0	+1	+2
f:	1	3	4	7	3	2

(MBA, BIT, Ranchi, 2000)

Solution. In order to determine the actual classes we should know two things: (*i*) the assumed mean, *i.e.*, working origin, and (*ii*) the common factor or the class-interval.

CALCULATING ASSUMED MEAN AND COMMON FACTOR

d	f	fd	fd^2
−3	1	−3	9
−2	3	−6	12
−1	4	−4	4
0	7	0	0
+1	3	+3	3
+2	2	+4	8
	$N = 20$	$\Sigma fd = -6$	$fd^2 = 36$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$13.1 = \sqrt{\frac{36}{20} - \left(\frac{-6}{20}\right)^2} \times i = \sqrt{1.8 - 0.09} \times i$$

$$13.1 = 1.31 \times i \quad \text{or} \quad i = \frac{13.1}{1.31} = 10$$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i \quad \text{or} \quad 107 = A - \frac{6}{20} \times 10 \quad \text{or} \quad A = 110.$$

Thus the working origin is 110 and the class-interval 10. Hence the various classes would be :

X:	75–85	85–95	95–105	105–115	115–125	125–135
f:	1	3	4	7	3	2

Illustration 20. The following data give the number of passengers travelling by Boeing 747 from one city to another in one week.

320, 290, 265, 300, 270, 200, 315

Calculate the mean and standard deviation and determine the percentage of cases that lie between (*i*) $\bar{X} \pm 1\sigma$, (*ii*) $\bar{X} \pm 2\sigma$, (*iii*) $\bar{X} \pm 3\sigma$. What percentage of cases lie outside 1σ, 2σ and 3σ limits ?

Solution. CALCULATION OF MEAN AND STANDARD DEVIATION

X	$(X - \bar{X})$	$(X - \bar{X})^2$
320	+40	1,600
290	+10	100
265	−15	225
300	+20	400
270	−10	100
200	−80	6,400
315	+35	1,225
$\Sigma X = 1960$	$\Sigma (X - \bar{X}) = 0$	$\Sigma (X - \bar{X})^2 = 10,050$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{1960}{7} = 280$$

$$\sigma = \sqrt{\frac{\Sigma (X - \bar{X})^2}{N}} = \sqrt{\frac{10050}{7}} = 37.89$$

(*i*) Cases lying between :

$$\bar{X} \pm 1\sigma = 280 \pm 37.89 = 242.11 \text{ to } 317.89.$$

There are 2 observations, *i.e.*, 200 and 320, that fall outside these limits. Hence the required percentage is 28.57.

(*ii*) Cases lying between :

$$\bar{X} \pm 2\sigma = 280 \pm 2 \times 37.89 = 204.22 \text{ to } 355.78.$$

There are only one observations that lies outside this limits. Hence the required percentage is 14.28.

(*iii*) Cases lying between :

$$\bar{X} \pm 3\sigma = 280 \pm 3 \times 37.89 = 166.33 \text{ to } 393.67.$$

Not a single observation lies outside this limit.

Illustration 21. A company has three establishments E_1, E_2 and E_3 in three cities. Analysis of the daily wages paid to the employees in the three establishments is given below :

	E_1	E_2	E_3
Number of employees	20	25	40
Average daily wage (Rs.)	305	300	340
Standard deviation (Rs.)	50	40	45

Find the average and the standard deviation of the wages of all the 85 employees in the company.

Solution.

$$\bar{X}_{123} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2 + N_3 \bar{X}_3}{N_1 + N_2 + N_3}$$

$$= \frac{(20 \times 305) + (25 \times 300) + (40 \times 340)}{20 + 25 + 40}$$

$$= \frac{6100 + 7500 + 13600}{85} = \frac{27200}{85} = 320$$

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

$$d_1 = |\bar{X}_1 - \bar{X}_{123}| = |305 - 320| = 15$$

$$d_2 = |\bar{X}_2 - \bar{X}_{123}| = |300 - 320| = 20$$

$$d_3 = |\bar{X}_3 - \bar{X}_{123}| = |340 - 320| = 20$$

$$\sigma_{123} = \sqrt{\frac{20(50)^2 + 25(40)^2 + 40(45)^2 + 20(15)^2 + 25(20)^2 + 40(20)^2}{20 + 25 + 40}}$$

$$= \sqrt{\frac{50000 + 40000 + 81000 + 4500 + 10000 + 16000}{85}}$$

$$= \sqrt{\frac{201500}{85}} = \sqrt{2370.59} = 48.69.$$

Thus the combined average wage is Rs. 320 and the combined standard deviation is Rs. 48.69.

Illustration 22. The mean and the standard deviation of a sample of size 10 were found to be 9.5 and 2.5 respectively. Later on, an additional observation became available. This was 15.0 and was included in the original sample. Find the mean and the standard deviation of the 11 observations.

Solution. $\bar{X} = \frac{\Sigma X}{N}$ or $9.5 \times 10 = \Sigma X$ or $\Sigma X = 95$

Adding 15, *i.e.*, 11th observation

$$\Sigma X = 95 + 15 = 110$$

$$\bar{X} = \frac{110}{11} = 10$$

$$\sigma^2 = \frac{\Sigma X^2}{N} - (\bar{X})^2 \quad \text{or} \quad (2.5)^2 = \frac{\Sigma X^2}{10} - (9.5)^2$$

or $$6.25 \times 10 = \Sigma X^2 - 902.5 \quad \text{or} \quad \Sigma X^2 = 965$$

On adding one more observation, *i.e.*, 15

$$\Sigma X^2 = 965 + (15)^2 = 1{,}190$$

$$\sigma^2 = \frac{\Sigma X^2}{N} - (\bar{X})^2 = \frac{1190}{11} - (10)^2$$

$$= 108.18 - 100 = 8.18 \quad \text{or} \quad \sigma = \sqrt{8.18} = 2.86.$$

Thus the mean and standard deviation of 11 observations are 10 and 2.86 respectively.

Illustration 23 A collar manufacturer is considering the production of a new style of collar to attract young men. The following statistics of neck circumference are available based on measurements of a typical group of the college students :

Mid-value (in inches) :	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0
Number of students :	2	16	36	60	76	37	18	3	2

Compute the standard deviation and use the criterion $\bar{X} \pm 3\sigma$ where σ is the standard deviation and $\bar{X}$ is the arithmetic mean to determine the largest and smallest size of the collar he should make in order to meet the needs of practically all the customers bearing in mind that collars are worn on average ½ inch longer than neck size. *(MBA, Delhi Univ, 2005)*

Solution. CALCULATION OF MEAN AND STANDARD DEVIATION

Mid-value in inches X	*No. of students* f	$(X-14)/10.5$ d	fd	fd^2
12.0	2	−4	−8	32
12.5	16	−3	−48	144
13.0	36	−2	−72	144
13.5	60	−1	−60	60
14.0	76	0	0	0
14.5	37	+1	+37	37
15.0	18	+2	+36	72
15.5	3	+3	+9	27
16.0	2	+4	+8	32
	$N = 250$		$\Sigma fd = -98$	$\Sigma fd^2 = 548$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 14 - \frac{98}{250} \times 0.5 = 13.8$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{548}{250} - \left(\frac{-98}{250}\right)^2} \times 0.5 = 1.43 \times 0.5 = 0.715$$

Largest and smallest neck size

$$= \bar{X} \pm 3\sigma = 13.8 \pm 3(0.715) = 11.655 \text{ and } 15.945.$$

Since collars are worn on an average 1/2 inch longer than the neck sizes, we should add 0.5 to these limits. Thus the smallest and largest sizes of collar should be

(11.655+0.5) and (15.945+0.5), *i.e.*,

12.155 or 12.2 inches and 16.445 or 16.4 inches.

Illustration 24. A study of the age of 100 persons grouped in intervals of 20–22, 22–24..... etc., revealed the mean age and standard deviation to be 32.02 and 13.18 respectively, While checking it was discovered that the observation 57 was misread as 27. Calculate the correct mean age and standard deviation. *(MBA, Delhi Univ., 1997)*

Solution. We are given : $N = 100$, $\bar{X} = 32.02$, $\sigma = 13.18$.

$$\bar{X} = \frac{\sum fX}{N} \text{ or } \sum fX = N\bar{X}$$

Uncorrected $\sum fX = 100 \times 32.02 = 3202$

$$\sigma^2 = \frac{1}{N}\sum fX^2 - (\bar{X})^2 \text{ or } \sum fX^2 = N[\sigma^2 + (\bar{X})^2]$$

Uncorrected $\sum fX^2 = 100\,[(13.18)^2 + (32.02)^2]$

$= 100(173.71 + 1025.28) = 100 \times 1198.99 = 119899$

Correct $\sum fX = 3202 - 27 + 57 = 3232$

Correct $\bar{X} = \frac{3232}{100} = 32.32$

Correct $\sum fX^2 = 119899 - (27) + (57)^2 = 119899 - 729 + 3249 = 122419$

Correct $\sigma^2 = \frac{\text{Correct } \sum fX^2}{N} - (\text{Correct Mean})^2$

Correct $\sigma^2 = \frac{122419}{100} - (32.32)^2 = 1224.19 - 1044.58 = 179.61$

Correct $\sigma = \sqrt{179.61} = 13.402$.

Illustration 25. For a group of 50 male workers, the mean and standard deviation of their daily wages are Rs. 63 and Rs. 9 respectively. For a group of 40 female workers, these are Rs. 54 and Rs. 6 respectively. Find the standard deviation of daily wages for the combined group of 90 workers. *(MBA, DU, 2002)*

Solution Combined Mean : $\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$

$N_1 = 50$, $\bar{X}_1 = 63$, $N_2 = 40$, $\bar{X}_2 = 54$

$$\bar{X}_{12} = \frac{(50 \times 63) + (40 \times 54)}{50 + 40} = \frac{3150 + 2160}{90} = \frac{5310}{90} = 59$$

Combined Standard Deviation

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$\sigma_1 = 9$, $\sigma_2 = 6$, $d_1 = |\bar{X}_1 - \bar{X}_{12}| = |63 - 59| = 4$, $d_2 = |\bar{X}_2 - \bar{X}_{12}| = |54 - 59| = 5$

$$\sigma_{12} = \sqrt{\frac{50(9)^2 + 40(6)^2 + 50(4)^2 + 40(5)^2}{50 + 40}}$$

$$= \sqrt{\frac{4050 + 1440 + 800 + 1000}{90}} = \sqrt{\frac{7290}{90}} = \sqrt{81} = 9$$

Illustration 26. Calculate variance and coefficient of variation from the following data :

Profits (Rs. crores)	*No. of Companies*
Less than 10	8
Less than 20	20
Less than 30	40
Less than 40	70
Less than 50	90
Less than 60	100

(MBA, Jodhpur Univ., 2001)

Solution. CALCULATION OF VARIANCE AND COEFFICIENT OF VARIATION

Profits (Rs. crores)	*No. of companies* f	*m.p.* X	$(X-35)/10$ d	fd	fd^2
0–10	8	5	–3	–24	72
10–20	12	15	–2	–24	48
20–30	20	25	–1	–20	20
30–40	30	35	0	0	0
40–50	20	45	+1	+20	20
50–60	10	55	+2	+20	40
	$N=10$			$\Sigma fd=-28$	$\Sigma fd^2=200$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 35 - \frac{28}{100} \times 10 = 35 - 2.8 = 32.2$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 10$$

$$= \sqrt{2-.0784} \times 10 = 1.3862 \times 10 = 13.862$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{13.862}{32.2} \times 100 = 43.05$$

$$\text{Variance} = \sigma^2 = (13.862)^2 = 192.155$$

Illustration 27. The prices of a Tea Company shares in Mumbai and Kolkata markets during the last ten months are recorded below:

Month	Mumbai	*Kolkata*	*Month*	Mumbai	*Kolkata*
January	105	108	June	127	125
February	120	117	July	109	125
March	115	120	August	110	120
April	118	130	September	104	110
May	130	100	October	112	135

Determine the Arithmetic Mean and Standard Deviation of the prices of shares. In which market are the share prices stable?

(MBA, HPU, 2002)

Solution. For determining in which market prices of shares are more stable, we shall compare the coefficient of variation. Let prices in Mumbai and Kolkata be denoted by X and Y respectively :

CALCULATION OF COEFFICIENT OF VARIATION

X	$(X-\bar{X})$ x	x^2	Y	$(Y-\bar{Y})$ y	y^2
105	−10	100	108	−11	121
120	+5	25	117	−2	4
115	0	0	120	+1	1
118	+3	9	130	+11	121
130	+15	225	100	−19	361
127	+12	144	125	+6	36
109	−6	36	125	+6	36
110	−5	25	120	+1	1
104	−11	121	110	−9	81
112	−3	9	135	+16	256
$\Sigma X = 1150$	$\Sigma x = 0$	$\Sigma x^2 = 694$	$\Sigma Y = 1190$	$\Sigma y = 0$	$\Sigma y^2 = 1018$

Mumbai : C.V. $= \frac{\sigma}{\bar{X}} \times 100$ Kolkata : C.V. $= \frac{\sigma}{\bar{Y}} \times 100$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{1150}{10} = 115 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{1190}{10} = 119$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{694}{10}} = 8.33 \qquad \sigma = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{1018}{10}} = 10.09$$

$$\text{C.V.} = \frac{8.33}{115} \times 100 = 7.24 \qquad \text{C.V.} = \frac{10.09}{119} \times 100 = 8.48$$

Since the coefficient of variation is less in Mumbai, hence the share price in Mumbai market shows greater stability.

Illustration 28. The following data give the number of finished articles turned out per day by different number of workers in a factory :

No. of articles :	18	19	20	21	22	23	24	25	26	27
No. of workers:	3	7	11	14	18	17	13	8	5	4

Find the mean, standard deviation and coefficient of variation of daily output of finished articles.

(MBA, Kurukshetra Univ.,2001)

Solution

CALCULATION OF $\bar{X}$, S.D. AND C.V.

No. of articles X	$(X-22)$ d	f	fd	fd^2
18	−4	3	−12	48
19	−3	7	−21	63
20	−2	11	−22	44
21	−1	14	−14	14
22	0	18	0	0
23	+1	17	+17	17
24	+2	13	+26	52
25	+3	8	+24	72
26	+4	5	+20	80
27	+5	4	+20	100
		$N = 100$	$\Sigma fd = 38$	$\Sigma fd^2 = 490$

$$\bar{X} = A + \frac{\Sigma fd}{N} = 22 + \frac{38}{100} = 22 + .38 = 22.38$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{490}{100} - \left(\frac{38}{100}\right)^2} = \sqrt{4.9 - .1444} = 2.18$$

$$\text{C.V.} = \frac{2.18}{22.38} \times 100 = 9.74$$

Illustration 29. Based on the frequency distribution given below, compute the following statistical measures to characterise the distribution.

(*i*) Coefficient of variation, (*ii*) Inter-quartile range, (*iii*) Modal value.

Annual Tax Paid (Rs. Thousand)	*No. of Managers*	*Annual Tax Paid (Rs. Thousand)*	*No. of Managers*
5–10	18	25–30	20
10–15	30	30–35	12
15–20	46	35–40	6
20–25	28		

(M.M.S, Bombay Univ., 2002)

Solution :

CALCULATION OF COEFFICIENT OF VARIATION, SEMI-INTER QUARTILE RANGE AND MODE

Annual Tax Paid (Rs. thousand)	*m.p. X*	*f*	*(X—22.5)/5 d*	*fd*	fd^2	*c.f.*
5–10	7.5	18	–3	–54	162	18
10–15	12.5	30	–2	–60	120	48
15–20	17.5	46	–1	–46	46	94
20–25	**22.5**	28	0	0	0	122
25–30	27.5	20	+1	+20	20	142
30–35	32.5	12	+2	+24	48	154
35–40	37.5	6	+3	+18	54	160
		N=160		$\Sigma fd = -98$	$\Sigma fd^2 = 450$	

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 22.5 - \frac{98}{160} \times 5 = 22.5 - 3.0625 = 19.4375$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{450}{160} - \left(\frac{-98}{160}\right)^2} \times 5$$

$$= \sqrt{2.8125 - 0.3752} \times 5 = 1.5612 \times 5 = 7.806$$

$$\text{Coefficient of Variation : C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{7.806}{19.4375} \times 100 = 40.1594$$

Semi-Inter Quartile Range :

$$\text{Semi-Inter Quartile Range} = Q_3 - Q_1$$

$$Q_1 = \text{Size of } \frac{N}{4} \text{ th observation} = \frac{160}{4} = 40\text{th observation}$$

Q_1 lies in the class 10–15.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 10 + \frac{40-18}{30} \times 5 = 10 + 3.67 = 13.67$$

$$Q_3 = \text{Size of } \frac{3N}{4} \text{ th observation} = \frac{3 \times 160}{4} = 120\text{th observation}$$

Q_3 lies in the class 20–25.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i = 20 + \frac{120-94}{28} \times 5 = 20 + 4.64 = 24.64$$

$$\text{Semi-Inter Quartile Range} = Q_3 - Q_1 = 24.64 - 13.67 = 10.97$$

Mode: Since the highest frequency is 46, the mode lies in the class 15–20,

$$\text{Mo} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 15 + \frac{16}{16+18} \times 5 = 15 + 2.35 = 17.35$$

Illustration 30. Calculate coefficient of variation from the following data :

Profits (Rs. crores):	10–20	20–30	30–40	40–50	50–60
No. of Cos. :	8	12	20	6	4

(MBA, Osmania Univ., 2002)

Solution. CALCULATION OF COEFFICIENT OF VARIATION

Profits (Rs. Crores)	*m.p. X*	*f*	*(X—35)/10 d*	*fd*	fd^2
10–20	15	8	–2	–16	32
20–30	25	12	–1	–12	12
30–40	35	20	0	0	0
40–50	45	6	+1	+6	6
50–60	55	4	+2	+8	16
		$N = 50$		$\Sigma fd = -14$	$\Sigma fd^2 = 66$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 35 - \frac{14}{50} \times 10 = 35 - 2.8 = 32.2$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{66}{50} - \left(\frac{-14}{50}\right)^2} \times 10$$

$$= \sqrt{1.32 - .0784} \times 10 = 11.14$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{11.14}{32.2} \times 100 = 34.6 \text{ per cent}$$

Illustration 31. Find the missing information from the following :

	Group I	*Group II*	*Group III*	*Combined*
Number	50	?	90	200
Standard Deviation	6	7	?	7.746
Mean	113	?	115	116

(MBA, HPU; MBA, Osmania Univ., 1997)

Solution. *Finding the number of observations in the second group.*

Let N_1, N_2, N_3 denote the number of observations in the 1st, 2nd and 3rd group respectively.

We are given $N_1 + N_2 + N_3 = 200$

$N_1 = 50$, $N_3 = 90$, $\therefore N_1 + N_3 = 140$

$N_2 = 200 - 140 = 60$

Finding Mean of the Second Group

$$\bar{X}_{123} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2 + N_3 \bar{X}_3}{N_1 + N_2 + N_3}$$

$\bar{X}_{123} = 116$, $N_1 + N_2 + N_3 = 200$, $\bar{X}_1 = 113$, $\bar{X}_3 = 115$

We have to find $\bar{X}_2$.

Substituting the given values

$$116 = \frac{50(113) + 60(\bar{X}_2) + 90(115)}{200}$$

$$116 \times 200 = 5650 + 60\ \bar{X}_2 + 10350$$

$$60\ \bar{X}_2 = 23200 - 16000 = 7200 \quad \text{or} \quad \bar{X}_2 = \frac{7200}{60} = 120$$

Finding S.D. of third group

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

$\sigma_{123} = 7.746, N_1 = 50, \sigma_1 = 6, N_2 = 60, \sigma_2 = 7, N_3 = 90$

$d_1 = |\bar{X}_1 - \bar{X}_{123}| = 113 - 116 = 3$

$d_2 = |\bar{X}_2 - \bar{X}_{123}| = 120 - 116 = 4$

$d_3 = |\bar{X}_3 - \bar{X}_{123}| = 115 - 116 = 1$

Substituting the values

$$7.746 = \sqrt{\frac{50(6)^2 + 60(7)^2 + 90\sigma_3^2 + 50(3)^2 + 60(4)^2 + 90(1)^2}{50 + 60 + 90}}$$

$$= \sqrt{\frac{1800 + 2940 + 90\sigma_3^2 + 450 + 960 + 90}{200}} = \sqrt{\frac{6240 + 90\sigma_3^2}{200}}$$

Squaring on both sides, we get

or $$(7.746)^2 = \frac{6240 + 90\sigma_3^2}{200}$$

or $$12{,}000 = 6240 + 90\sigma_3^2 \text{ or } 90\sigma_3^2 = 12{,}000 - 6240$$

$$\sigma_3^2 = \frac{5760}{90} = 64 \quad \text{or} \quad \sigma_3 = \sqrt{64} = 8$$

Thus the missing values are :

$N_2 = 60, \ \bar{X}_2 = 120, \sigma_3 = 8.$

Illustration 32. Given below are the daily wages, in rupees, of 60 workers in a factory manufacturing plastic products :

23	48	51	64	72	82	56	33	50	42
35	88	77	65	39	52	48	64	49	57
41	73	62	49	32	54	67	46	55	50
82	44	75	56	51	63	59	69	53	42
75	85	68	65	52	45	42	57	20	57
46	51	20	16	62	46	54	40	55	71

(*a*) Form a frequency distribution, taking the lowest class-interval as 10–20.

(*b*) Calculate the Standard Deviation and Coefficient of Variation of this distribution.

(MBA, HPU,2002)

Solution. FORMATION OF FREQUENCY DISTRIBUTION
CALCULATION OF COEFFICIENT OF VARIATION

Wages (Rs.)	*Tally*	*Frequency f*	*m.p. X*	*(X—45)/10 d*	*fd*	fd^2
10–20	|	1	15	−3	−3	9
20–30	||	2	25	−2	−4	8
30–40	||||	4	35	−1	−4	4
40–50	𝍸 𝍸 |||	13	45	0	0	0
50–60	𝍸 𝍸 𝍸 𝍸 |	21	55	+1	+21	21
60–70	𝍸 ||||	9	65	+2	+18	36
70–80	|||| |	6	75	+3	+18	54
80–90	||||	4	85	+4	+16	64
		$N = 60$			$\Sigma fd = 62$	$\Sigma fd^2 = 196$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 45 + \frac{62}{60} \times 10 = 45 + 10.33 = 55.33$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{196}{60} - \left(\frac{62}{60}\right)^2} \times 10$$

$$= \sqrt{3.267 - 1.068} \times 10 = 1.483 \times 10 = 14.83$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{14.83}{55.33} \times 100 = 26.8 \text{ per cent.}$$

Illustration 33. Particulars regarding the income of two villages are given below :

	Village X	*Village Y*
Number of employees	600	500
Average income (in Rs.)	1750	1860
S.D. of income (in Rs.)	100	81

(*i*) In which village is the variation in income greater ?

(*ii*) What is the combined standard deviation of the village *X* and village *Y* put together ?

Solution. (*i*) Compare coefficient of variation to find out in which village there is greater variation in income.

$$\text{Village } X : \text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{100}{1750} \times 100 = 5.71$$

$$\text{Village } Y : \text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{81}{1860} \times 100 = 4.35$$

Since coefficient of variation is more in case of village *X*, hence in this village there is greater variation in the income of employees.

(*ii*) For finding combined variation we have to find first the combined mean.

$$\bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} = \frac{(600 \times 1750) + (500 \times 1860)}{600 + 500} = \frac{1050000 + 930000}{1100} = 1800$$

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$$d_1 = |\bar{X}_1 - \bar{X}_{12}| = |1750 - 1800| = 50$$

$$d_2 = |\bar{X}_2 - \bar{X}_{12}| = |1860 - 1800| = 60$$

$$\sigma_{12} = \sqrt{\frac{600(100)^2 + 500(81)^2 + 600(50)^2 + 500(60)^2}{600 + 500}}$$

$$= \sqrt{\frac{6000000 + 3280500 + 1500000 + 1800000}{1100}}$$

$$= \sqrt{\frac{12580500}{1100}} = \sqrt{11436.82} = 106.94.$$

Illustration 34. The profits (in Rs. lakhs) earned by 100 companies during 2009-10 are shown below :

Profits (Rs. lakhs)	*No. of Companies*	*Profits (Rs. lakhs)*	*No. of Companies*
20–30	4	60–70	15
30–40	8	70–80	10
40–50	18	80–90	8
50–60	30	90–100	7

Compute (*a*) Mean, (*b*) Median and (*c*) Standard Deviation.

Solution. COMPUTATION OF MEAN, MEDIAN AND STANDARD DEVIATION

Profits (Rs. lakhs)	*No. of cos. f*	*m.p. X*	*(X—55)/10 d*	*fd*	fd^2	*c.f.*
20–30	4	25	–3	–12	36	4
30–40	8	35	–2	–16	32	12
40–50	18	45	–1	–18	18	30
50–60	30	**55**	0	0	0	60
60–70	15	65	+1	+15	15	75
70–80	10	75	+2	+20	40	85
80–90	8	85	+3	+24	72	93
90–100	7	95	+4	+28	112	100
	$N = 100$			$\Sigma fd = 41$	$\Sigma fd^2 = 325$	

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 55 + \frac{41}{100} \times 10 = 55 + 4.1 = 59.1$$

$$\text{Med.} = \text{Size of } \frac{N}{2} \text{ th observation} = \frac{100}{2} = 50\text{th observation}$$

Median lies in the class 50–60.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i = 50 + \frac{50-30}{30} \times 10 = 56.67$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{325}{100} - \left(\frac{41}{100}\right)^2} \times 10$$

$$= \sqrt{3.25 - .1681} \times 10 = 1.756 \times 10 = 17.56$$

Illustration 35. Find the mean, median and standard deviation of the weight of billet in 8 gms as given in the following table :

(Wt. in gms)	*Frequency*	*(Wt. in gms)*	*Frequency*
210–215	8	230–235	14
215–220	13	235–240	10
220–225	16	240–245	7
225–230	29	245–250	3

Solution : CALCULATION OF $\bar{X}$, MED. AND STANDARD DEVIATION

Wt. in gms	*f*	*m.p. X*	*(X–227.5/5) d*	*fd*	fd^2	*c.f.*
210–215	8	212.5	–3	–24	72	8
215–220	13	217.5	–2	–26	52	21
220–225	16	222.5	–1	–16	16	37
225–230	29	**227.5**	0	0	0	66
230–235	14	232.5	+1	+14	14	80
235–240	10	237.5	+2	+20	40	90
240–245	7	242.5	+3	+21	63	97
245–250	3	247.5	+4	+12	48	100
	$N=100$			$\Sigma fd = +1$	$\Sigma fd^2=305$	

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 227.5 + \frac{1}{100} \times 5 = 227.55$$

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i$$

$$\text{Med.} = \text{Size of } \frac{N}{2}\text{th observation} = \frac{100}{2} = 50\text{th observation}$$

Median lies in the class 225–230.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i = 227.5 + \frac{50-37}{29} \times 5 = 227.5 + 2.24 = 229.74$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{3.05 - .0001} \times 5 = 1.7464 \times 5 = 8.732$$

Illustration 36. The following distribution shows the turnover of the branches of a group of multiple-shops in March 2010:

Turnover (in Rs. lakhs)	*No. of Shops*
5 and under 10	8
10 and under 15	18
15 and under 20	42
20 and under 25	62
25 and under 30	30
30 and under 35	10
35 and over	4

Using assumed mean of Rs. 22,500, calculate (*i*) Mean, (*ii*) Standard deviation, and (*iii*) Coefficient of variation.

Solution. CALCULATION OF MEAN, STANDARD DEVIATION AND COEFFICIENT OF VARIATION

Turnover (in Rs. lakhs)	*m.p.* X	f	$(X-22.5)/5$ d	fd	fd^2
5–10	7.5	8	–3	–24	72
10–15	12.5	18	–2	–36	72
15–20	17.5	42	–1	–42	42
20–25	**22.5**	62	0	0	0
25–30	27.5	30	+1	+30	30
30–35	32.5	10	+2	+20	40
35–40	37.5	4	+3	+12	36
		$N = 174$		$\Sigma fd = -40$	$\Sigma fd^2 = 292$

Mean: $$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 22.5 - \frac{40}{174} \times 5 = 22.5 - 1.15 = 21.35$$

Standard Deviation : $$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{292}{174} - \left(\frac{40}{174}\right)^2} \times 5$$

$$= \sqrt{1.678 - .053} \times 5 = 1.275 \times 5 = 6.375$$

Coefficient of Variation : $\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{6.375}{21.35} \times 100 = 29.86$ per cent.

Illustration 37. For two firms *A* and *B* belonging to same industry, the following details are available :

	Firm *A*	Firm *B*
Number of Employees :	100	200
Average monthly wage :	Rs. 4,800	Rs. 5,100
Standard deviation :	Rs. 600	Rs. 540

Find (*i*) Which firm pays out larger amount as wages ?

(*ii*) Which firm shows greater variability in the distribution of wages ?

(*iii*) Find average monthly wage and the standard deviation of the wages of all employees in both the firms.

Solution. (*i*) For finding out which firm pays larger amount, we have to find out ΣX. *(MBA, Delhi Univ., 2006)*

$$\bar{X} = \frac{\Sigma X}{N} \quad \text{or} \quad \Sigma X = N\bar{X}$$

Firm *A* : $N = 100$, $\bar{X} = 4800$, $\therefore \Sigma X = 100 \times 4800 = 4,80,000$

Firm *B* : $N = 200$, $\bar{X} = 5100$, $\therefore \Sigma X = 200 \times 5100 = 10,20,000$

Hence firm *B* pays larger amount as monthly wages.

(*ii*) For finding out which firm shows greater variability in the distribution of wages, we have to calculate coefficient of variation.

$$\text{Firm } A: \ C.V. = \frac{\sigma}{\bar{X}} \times 100 = \frac{600}{4800} \times 100 = 12.50.$$

$$\text{Firm } B: C.V. = \frac{\sigma}{\bar{X}} \times 100 = \frac{540}{5700} \times 100 = 10.59.$$

Since coefficient of variation is greater in case of firm *A*, hence it shows greater variability in the distribution of wages.

(*iii*) Combined average monthly wage :

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$N_1 = 100, \ \bar{X}_1 = 4800, N_2 = 200, \ \bar{X}_2 = 5100$$

$$\text{Hence } \bar{X}_{12} = \frac{(100)(4800) + (200)(5100)}{100 + 200} = \frac{480000 + 1020000}{300} = \text{Rs. } 5{,}000$$

Combined Standard Deviation

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$$N_1 = 100, \sigma_1 = 600, N_2 = 200, \sigma_2 = 540$$

$$d_1 = |\bar{X}_1 - \bar{X}_{12}| = |4800 - 5000| = 200$$

$$d_2 = |\bar{X}_2 - \bar{X}_{12}| = |5100 - 5000| = 100$$

$$\sigma_{12} = \sqrt{\frac{100(600)^2 + 200(540)^2 + 100(200)^2 + 200(100)^2}{100 + 200}}$$

$$= \sqrt{\frac{36000000 + 58320000 + 4000000 + 2000000}{300}}$$

$$= \sqrt{\frac{100{,}320{,}000}{300}} = 578.27$$

Hence the combined standard deviation is Rs. 578.27.

Illustration 38. From the following frequency distribution of heights of 360 boys in the age-group 15–20 years, calculate the: (*i*) arithmetic mean ; (*ii*) coefficient of variation ; and (*iii*) quartile deviation.

Heights (in cms)	*No. of boys*	*Height (in cms)*	*No. of boys*
126–130	31	146–150	60
131–135	44	151–155	55
136–140	48	156–160	43
141–145	51	161–165	28

Solution. CALCULATION OF $\bar{X}$, Q.D., AND C.V.

Height (in cms)	*m.p. X*	*f*	*(X–143)/5 d*	*fd*	fd^2	*c.f.*
126–130	128	31	–3	–93	279	31
131–135	133	44	–2	–88	176	75
136–140	138	48	–1	–48	48	123
141–145	143	51	0	0	0	174
146–150	148	60	+1	+60	60	234
151–155	153	55	+2	+110	220	289
156–160	158	43	+3	+129	387	332
161–165	163	28	+4	+112	448	360
		N =360		Σfd= 182	Σfd^2=1618	

(*i*) $$\bar{X} = A + \frac{\Sigma fd}{N} \times i$$

$$= 143 + \frac{182}{360} \times 5 = 143 + 2.53 = 145.53$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{1618}{360} - \left(\frac{182}{360}\right)^2} \times 5$$

$$= \sqrt{4.494 - 0.256} \times 5 = 2.059 \times 5 = 10.295$$

(*ii*) $$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{10.295}{145.53} \times 100 = 7.074 \text{ per cent}$$

(*iii*) $$Q.D. = \frac{Q_3 - Q_1}{2}$$

Q_1 = Size of $\frac{N}{4}$ th observation = $\frac{360}{4}$ = 90th observation

Q_1 lies in the class 136–140. But the real limit of this class is 135.5–140.5.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 135.5 + \frac{90 - 75}{48} \times 5 = 135.5 + 1.56 = 137.06$$

Q_3 = Size of $\frac{3N}{4}$ th observation = $3 \times \frac{360}{4}$ = 270th observation

Q_3 lies in the class 151–155. But the real limit of this class is 150.5–155.5.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i = 150.5 + \frac{270 - 234}{55} \times 5 = 150.5 + 3.27 = 153.77$$

$$Q.D. = \frac{153.77 - 137.06}{2} = 8.355.$$

Illustration 39. A welfare organisation introduced on education scholarship scheme for the school going children of a backward village. The rates of scholarship were fixed as given below :

Age Group (in yrs.)	*Amount of Scholarship per month (Rs.)*
5–7	300
8–10	400
11–13	500
14–16	600
17–19	700

The ages (years) of 30 school going children are noted as 11, 8, 10, 5, 7, 12, 7, 17, 5, 13, 9, 8, 10, 15, 7, 12, 6, 7, 8, 11, 14, 18, 6, 13, 9, 10, 6, 15, 3, 5 years respectively. Calculate mean and standard deviation of monthly scholarship. Find out the total monthly scholarship amount being paid to the students. (*MBA, IGNOU, 2002*)

Solution : Let us first classify the given data in various class intervals.

Age group (yrs)	*Tally bars*	*Freq.*
5–7	𝍸 𝍸	10
8–10	𝍸 |||	8
11–13	𝍸 ||	7
14–16	|||	3
17–19	||	2

CALCULATION OF MEAN AND STANDARD DEVIATION

Age group (yrs.)	*m.p.* m	f	$(m-12)/3$ d	fd	fd^2
5–7	6	10	−2	−20	40
8–10	9	8	−1	−8	8
11–13	12	7	0	0	0
14–16	15	3	+1	+3	3
17–19	18	2	+2	+4	8
		$N = 30$		$\Sigma fd = -21$	$\Sigma fd^2 = 59$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 12 - \frac{21}{30} \times 3 = 12 - 2.1 = 9.9$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2} \times i = \sqrt{\frac{59}{30} - \left(\frac{-21}{30}\right)^2} \times 3$$

$$= \sqrt{1.967 - 0.49 \times 3} = 1.2153 \times 3 = 3.65$$

CALCULATION OF TOTAL MONTHLY SCHOLARSHIP

No. of students	*Amount of scholarship per month (Rs.)*	*Total Monthly Scholarship (Rs.)*
10	300	3000
8	400	3200
7	500	3500
3	600	1800
2	700	1400
		Rs. 12900

Illustration 40. The performance of two teams is summarized below :

	Mean	*S.D.*
Team '*A*'	10	2
Team '*B*'	15	4

Which is more consistent team? *(MBA, G.G.S.I.P. Univ., 2009)*

Solution : For finding out which is more consistent team, we compare the coefficient of variation of the two teams.

Team *A* $\quad C.V. = \frac{\sigma}{\bar{X}} \times 100 = \frac{2}{10} \times 100 = 20$

Team *B* $\quad C.V. = \frac{\sigma}{\bar{X}} \times 100 = \frac{4}{15} \times 100 = 26.67$

Since coefficient of variation is less for Team '*A*', hence Team '*A*' is more consistent.

PROBLEMS

1-A : Answer the following questions, each question carries **one** mark :

(*i*) What is the formula for coefficient of variation ?

(*ii*) What is range ?

(*iii*) How quartile deviation is calculated ?

(*iv*) What is interquartile range ?

(*v*) State the formula for standard deviation. *(MBA., Madurai-Kamaraj Univ., 2003)*

(*vi*) What is Lorenz Curve ?

(*vii*) Is standard deviation independent of change of scale and origin .

(*viii*) Is the sum of deviations from mean is always least ?

(*ix*) Is variance the square of standard deviation ?

(*x*) Give the formula for combined standard deviation of two sets of data.

1-B : Answer the following questions, each question carries **four** marks :

(*i*) What are the various methods of measuring variation ?
(*ii*) Distinguish between mean deviation and standard deviation.
(*iii*) What are the properties of a good measure of variation ?
(*iv*) Distinguish between absolute and relative measures of dispersion.
(*v*) Why standard deviation is most widely used as a measure of variations.
(*vi*) What are the uses of Lorenz Curve ?

2. (*a*) Explain the term variation. What purpose does a measure of variation serve ? In the light of these, comment on some of the well-known measures of variation along with their respective merits and demerits.
(*b*) Point out the difference between absolute and relative variation.
(*c*) Under what circumstances range is more meaningful than any other measure of variation ? *(MBA, HPU, 2009)*

3. (*a*) What are the requisites of a good measure of dispersion ? Why is the standard deviation usually chosen as a measure of variation ?
(*b*) Explain how measure of central tendency and measure of variation complement each other in describing mass of data.

4. What is coefficient of variation ? What purpose does it serve ? Also distinguish between 'variance' and 'coefficient of variation'.

5. What do you understand by standard deviation? Explain its usefulness. Highlight its important properties.

6. With an example show that mean is dependent on both origin and scale while standard deviation is dependent on scale but not on origin.

7. (*a*) State the different measures of central tendency and variation.
(*b*) What do you understand by "coefficient of variation" ? Discuss its importance in business problems.

8. What is Lorenz Curve ? How is it drawn ? In what way does it help in studying variations of two or more distributions ? Illustrate with the help of an example.

9. (*a*) Explain with suitable examples the term 'variation'. Mention some common measures of variation and describe the one which you think is the most important.
(*b*) Critically examine the different methods of measuring variation. Which of these do you consider as the best and why ? *(MBA, KU, 2002)*

10. Explain and illustrate how the measures of variation afford a supplement to the information about frequency distribution furnished by averages. *(MBA, Delhi Univ., 2001)*

11. Explain briefly what is meant by 'Variation' of data. State the requisites of a satisfactory measure of dispersion and examine in their light any two measures of dispersion.

12. What are the uses of coefficient of variation in statistical analysis ?

13. Describe the various methods of measuring variation used in business decisions pointing out their limitations, if any. *(MBA, Delhi Univ., 2004)*

14. (*a*) Describe briefly the various methods of measuring variation in health statistics.
(*b*) What are the uses of coefficient of variation in statistical analysis ?

15. (*a*) Briefly describe the characteristics of a standard deviation as a measure of variation.
(*b*) Comment on the following statements :
(*i*) If the mean and standard deviation of n observations $x_1, x_2, \ldots\ldots x_n$ be $\overline{X}$ and σ respectively then the mean and the standard deviation of $-x_1, -x_2, \ldots\ldots -x_n$ will be $\overline{X}$ and $-\sigma$ respectively.
(*ii*) "After settlement the average daily wage in a factory had increased from Rs. 85 to Rs. 90 and the standard deviation had increased from 10 to 12.5. After settlement the wage has become higher and more uniform."

16. (*a*) Ten observations have mean 20 and standard deviation 5. If each of these 10 observations is doubled then the standard deviation of new observations will be...
(*b*) State whether the following statement is *true* or *false*:
Range is a measure of variation which gives us information about scatter of values about the measure of a central tendency.

17. Calculate mean deviation for the following frequency distribution :

No. of colds experienced in 12 months	*No. of persons*	*No. of colds experienced in 12 months*	*No. of persons*
0	15	5	95
1	46	6	82
2	91	7	26
3	162	8	13
4	110	9	2

[M.D.=1.47]

18. Calculate the appropriate measure of variation from the following data:

Daily Wages (in Rs.)	*No. of wage earners*	*Daily Wages (in Rs.)*	*No. of wage earners*
Less than 85	14	91–93	18
85–87	62	Over 93	7
88–90	99		

19. Compute coefficient of variation from the following table :

Weekly Income (in Rs.)	*No. of employees*	*Weekly Income (in Rs.)*	*No. of employees*
1300–1399	30	1700–1799	60
1400–1499	46	1800–1899	50
1500–1599	58	1900–1999	20
1600–1699	76		

20. The following table gives the fluctuations in the prices of shares of two companies *A* and *B*. Find out which of them shows greater variability. Comment on the result.

Price (in Rs.) Share A	*Price (in Rs.) Share B*	*Price (in Rs.) Share A*	*Price (in Rs.) Share B*
318	2542	324	2545
322	2522	315	2530
325	2534	308	2556
312	2532	319	2530

[Share *A* : C.V. = 1.75 ; Share *B* : C.V. = 0.42]

21. From an analysis of monthly wages paid to workers in two organisations *C* and *D*, the following results were obtained :

	C	*D*
No. of workers	550	600
Average monthly wages (Rs.)	2260	2348
Variance of the distribution of wages	100	144

Obtain the average monthly wages and the variability in individual wages of all the workers in the organisations taken together.

22. From the following table giving data regarding income of workers in two factories draw a graph (Lorenz curve) to show which factory has greater inequalities of income :

Income (Rs.)	*Factory A*	*Factory B*
Below 4500	6,000	5,000
4,500—5,000	4,250	4,500
5,000—6,000	3,600	4,800
6,000—7,000	1,500	2,200
7,000—8,000	650	1,500

[Factory *A*]

23. A distribution consists of three parts, characterised as follows :

Parts	*Number of items*	*Arithmetic average*	*S.D.*
1	200	5	3
2	250	10	4
3	300	15	5

Show that the arithmetic average of the whole distribution is 10.67 and its standard deviation is 5.83 approximately.

24. Lives of two models of refrigerators in a recent survey are :

Life (No. of years)	*No. of refrigerators 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 of these refrigerators ? Which model has greater uniformity ? *(MBA, Bharthidasan Univ., 2001; IAS, 2002; C.S.E., 2002)*

[Model A: $\bar{X} = 5.12$, C.V. = 54.9; Model B: $\bar{X} = 6.16$, C.V. = 36.2]

25. The mean of two samples of size 50 and 100 respectively is 54.1 and 50.3 and the standard deviations are 8 and 7. Obtain the standard deviation of the sample of size 150 obtained by combining the two samples.

[7.55]

26. 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 item 8 was incorrect. Calculate the mean and standard deviation if (*i*) the wrong item is omitted, and (*ii*) it is replaced by12. [10.1, 2.26, 10.2,2]

27. The mean and standard deviation of 1,000 observations of a frequency distribution (grouped in intervals 0–10, 10–20, etc.) were found to be 35 and 8. Later it was discovered that in calculating these values the errata which was supplied with the data was not considered. The errata read as follows :

Group 0–10 for frequency 25 read frequency 52
Group 10–20 for frequency 75 read frequency 57
Group 20–30 for frequency 121 read frequency 21
Group 30–40 for frequency 137 read frequency 73
Group 40–50 for frequency 59 read frequency 95.

Calculate the correct mean and standard deviation.

28. An analysis of the weekly wages paid to workers in two firms *A* and *B* belonging to the same industry, gives the following results:

	Firm A	*Firm B*
Number of wage-earners	550	650
Average daily wages	100	95
Standard deviation	$\sqrt{90}$	$\sqrt{120}$

(*a*) Which firm *A* or *B* pays out large amount as daily wages ?

(*b*) In which firm *A* or *B* is there greater variability in individual wages ?

(*c*) What are the measures of (*i*) average daily wages and (*ii*) standard deviation in the distribution of individual wages of all workers in the two firms taken together ? *(MBA, M.D. Univ.; Diploma in Mgt., AIMA, 1999)*

29. A factory produces two types of electric lamps *A* and *B*. In an experiment relating to their life, the following results were obtained :

Length of life (in hours)	No. of lamps *A*	No. of Lamps *B*
500–700	5	4
700–900	11	30
900–1,100	26	12
1,100–1,300	10	8
1,300–1,500	8	6

Compare the variability of the life of the two types using coefficient of variation.
[$CV_A = 21.64$; $CV_B = 23.41$] *(AIMA, 2005)*

30. In a small town, a survey was conducted in respect of profits made by retail shops. The following results were obtained:

Profit or Loss (in '000 Rs.)	*No. of shops*	*Profit or Loss (in '000 Rs.)*	*No. of shops*
–4 to –3	4	1 to 2	56
–3 to –2	10	2 to 3	40
–2 to –1	22	3 to 4	24
–1 to 0	28	4 to 5	18
0 to 1	38	5 to 6	10

Calculate (*i*) the average profit made by a retail shop.
(*ii*) total profit made by all shops, and (*iii*) the coefficient of variation of earnings.
[(*i*) 1348; (*ii*) 3,37,000; (*iii*) 152.8)]

31. The following is the distribution of amounts spent for research and development and for marketing for the year 2010 by 10 drug firms and cosmetic firms :

Expenditure (Lakhs of rupees)	*Drug Cos.* *R and D*	*Marketing*	*R and D*	*Cosmetic Cos.* *Marketing*
10–20	2	3	5	0
20–30	2	2	3	2
30–40	3	3	2	4
40–50	3	2	0	4

(*a*) Compute the arithmetic mean for each type of company, for each type of expenditure.

(*b*) Compute the standard deviation for each type of company, for each type of expenditure.

(*c*) What conclusion do you draw from the results ? *(MBA, Kurukshetra Univ.)*

[$\overline{X}$ = 32, 29, 22, 37; σ = 11, 11.14, 7.81, 7.48]

32. The life of two types of tyres in a sample survey is given below :

Life (in km)	*Type A*	*Type B*
5,000–10,000	18	15
10,000–15,000	22	24
15,000–20,000	26	30
20,000–25,000	25	18
25,000–30,000	9	13

(*a*) Which of the two types of tyre give a higher average life?

(*b*) If prices are same for both the types, which type would you prefer and why ? *(MBA, Delhi Univ., 1999)*

[(*a*) type *B*, (*b*) Type *B*]

33. A collar manufacturer is considering the production of new style of collar to attract young men. The following statistics of neck circumference are available based on measurement of a typical group:

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 arithmetic mean and standard deviation and comment on the results.

[14.22, 0.70]

34. Calculate the arithmetic mean, median and standard deviation for the following distribution :

Height (inches)	*No. of persons*	*Height (inches)*	*No. of persons*
60 less than 63	4	69 less than 72	33
63 less than 66	14	72 less than 75	8
66 less than 69	59	75 less than 78	2

[68.32; 68.14; 2.79]

35. The following data relate to the number of bonds applied for number of applicants and number of bonds allotted to each applicant by the Rayon Silk Mfg. (Wvg.) Co. Ltd. in Jan. 2004:

No. of Bonds applied for	*No. of Applicants*	*No. of bonds allotted to each applicant*
5–15	61,685	5
20–45	18,879	10
50–105	7,230	15
110–185	647	20
190–300	177	25
305–325	79	30

Calculate the average number of bonds allotted to each applicant and the standard deviation.

36. The standard deviation of a distribution of 100 values was Rs. 2. If the sum of the squares of the actual values was Rs. 3,600, what was the mean of this distribution ?

[5.66]

37. 32 trials of a process to finish a certain job revealed the following information :

Mean time taken to complete the job = 1 hr. 20 mts.

Standard deviation =16mts.

Another set of 8 trials gave mean time as 100 minutes and standard deviation 25 minutes.

Find the combined mean and standard deviation.

[$\overline{X}_{12}$ = 84, σ_{12} =19.84]

38. The following table relates to the profits and losses of 100 firms. Calculate the average profits and the standard deviation of profits :

Profits & Loss	*No. of firms*	*Profits & Loss*	*No. of firms*
5,000–6,000	8	0–1,000	6
4,000–5,000	12	(–)1,000–0	5
3,000–4,000	30	(–)2,000–(–)1,000	8
2,000–3,000	10	(–)3,000–(–)2,000	9
1,000–2,000	5	(–)4,000–(–)3,000	7

[$\bar{X} = 531$, $\sigma = 214.8$, C.V.= 40.5%]

39. A study of 241 authors revealed the following data on the distribution of age :

Age (years)	*Number of Authors*
up to 30	20
up to 40	73
up to 50	80
up to 60	44
up to 70	22
up to 80	2

Compute the mean and coefficient of variation of the distribution.

[$\bar{X} = 44.21$, $\sigma = 11.18$, C.V. = 25.29]

40. A survey of domestic consumption of electricity gave the following distribution of the no. of units consumed:

Number of units	*Number of consumers*	*Number of units*	*Number of consumers*
0–200	9	800–1000	45
200–400	18	1000–1500	38
400–600	27	1500–2000	20
600–800	32	2000 and above	2

Use a graphical method to calculate as accurately as possible the two quartiles and hence find the quartile deviation.

[$Q_1 = 570$, $Q_3 = 1250$, $Q.D. = 340$]

41. The Shareholder Research Bureau of India conducted recently a research study on the price behaviour of three leading industrial shares *A*, *B*, *C* for the period 2005 to 2008, the results of which are published as following in the quarterly journal :

Share	*Average Price*	*Standard Deviation*	*Current selling price*
	(Rs.)	*(Rs.)*	*(Rs.)*
A	18.00	5.40	36.00
B	22.50	4.50	34.75
C	24.00	6.00	39.00

(*i*) Which share in your opinion appears to be more stable in value ?

(*ii*) If you are the holder of all three shares which one would you like to dispose of at present, and why ?

[C.V. : (A) = 30, (B) = 20, (C) = 25]

42. Find the missing value from the following table :

Sub-group	*N*	$\bar{X}$	*Variance*
A	?	25	9
B	250	?	96
C	300	15	?
	750	16	51.733

[$N_3 = 200$, $\bar{X}_2 = 10$, $\sigma^2_3 = 25$]

43. The number examined, the mean weight and the standard deviation in each group of examination and two medical examiners are given below. Find the weight and standard deviation of both groups taken together.

Medical Examiner	*Number Examined*	*Mean Weight*	*Standard Deviation*
A	50	113 pounds	6.5 pounds
B	60	120 pounds	8.2 pounds

44. Blood serum cholesterol levels of 10 persons are as under :

240, 260, 290, 245, 255, 288, 272, 263, 277, 250.

Calculate standard deviation with the help of assumed mean.

[$\sigma = 16.48$]

45. Mean and standard deviation of the following continuous series are 31 and 15.94 respectively. The distribution after step deviations is as follows :

d :	–3	–2	–1	0	1	2	3
f :	10	15	25	25	10	11	5

Determine the actual class-intervals. *(MBA, Vikram Univ., 2002)*

[$i = 10$, $A = 35$, 0–10, 10–20, 20–30, 30–40 etc.]

46. The mean of 5 observations is 15 and the variance is 9. If two more observation having values –3 and 10 are combined with these 5 observations, what will be the mean and variance of 7 observations ?

47. Daily rated employees in SLM-Devilal, an engineering firm, earn Rs. 66 per day. The workers estimate that on the average they turn out 30 pieces per day with a standard deviation of six pieces per day. Under a suggested piece rate plan how much will they ask per piece if they wish to earn more than their present daily income 90 per cent of the time ?

48. The coefficients of variation of wages of male workers and female workers are 55 per cent and 70 per cent respectively, while the standard deviations are 220 and 15.4 respectively. Calculate the overall average wage of all workers given that 80 per cent of the workers are male.

49. Calculate the standard deviation from the following data :

Temperature (C)	*No. of days*	*Temperature (C)*	*No. of days*
–40 to –30	10	0 to 10	65
–30 to –20	24	10 to 20	180
–20 to –10	30	20 to 30	14
–10 to –0	42		

(MBA, Jodhpur Univ., 1996)

50. Calculate coefficient of variation from the following data :

No. of days absent:	0–5	5–10	10–15	15–20	20–25	25–30
No. of Students :	29	140	250	108	52	21

[C.V. = 41.32%]

51. The index number of prices of Cotton and Coal shares in April 2008 were as under :

Month	*Jan.*	*Feb.*	*March*	*April*	*May*	*June*	*July*	*Aug.*	*Sept.*	*Oct.*
Cotton	188	178	173	164	172	184	184	185	211	217
Coal	131	130	130	129	129	120	127	127	130	137

Which of the shares you consider more variable in price?

52. (*a*) The mean and standard deviation of 17 observations were found to be 25 and 5 respectively. Later on it was found that two values 51 and 31 were wrongly read as 35 and 13 respectively. Find the correct mean and standard deviation.

[Correct : $\overline{X} = 27$, $\sigma = 6.96$]

(*b*) In a survey, data on daily wages paid to workers of two factories *A* and *B* are as follows :

(No. of workers)

Daily wages (Rs.) :	*20 – 30*	*30 – 40*	*40 – 50*	*50 – 60*	*60 – 70*	*70 – 80*	*80 – 90*
Factory *A* :	15	30	44	60	30	14	7
Factory *B* :	25	40	60	35	20	15	5

Find out :

(*a*) Which factory pays higher average wages ? By how much ?

(*b*) Wages of which have greater variability.

(*c*) Monthly wage bill of both Factories (month = 25 days).

53. The number of employees, wages per employee and the variance of the wages per employee for two factories is given below:

	Factory A	*Factory B*
No. of employees	100	150
Average wage per employee per month (Rs.)	3,200	2,800
Variance of the wages per employee per month (Rs.)	625	729

(*a*) In which factory is there greater variation in the distribution of wages per employee ?

(*b*) Suppose in factory *B*, the wages of an employee were wrongly noted as Rs. 3050 instead of 3650, what would be the correct variance for factory *B* ? (*MBA, Kumaun Univ., 2003*)

54. Name the various measures of dispersion. How would you compare the performance of two companies which reported profits for last five years as follows :

Company I	:	4.0	4.1	4.3	4.0	4.1
Company II	:	7.3	–3.7	8.4	–2.5	11.0

(*MBA, M.D. Univ., 2000*)

55. Calculate variance and coefficient of variation from the following data :

Profits (*Rs. crore*)	*No. of Cos.*	*Profits* (*Rs. crore*)	*No. of Cos.*
Less than 10	8	Less than 40	70
" " 20	20	" " 50	90
" " 30	40	" " 68	100

(*MBA, Guru Jameshwar Univ., 2003*)

56. Assume that your pathology lab. provides the following details of blood test carried out on 100 patients for diabetes :

Blood Sugar	: 90 –100	100 –110	110 –120	120 –130	130 –140	140 –150	150 –160
No. of Patients	: 16	20	22	25	10	4	3

Calculate coefficient of variation. What inference do you draw. (*MBA, HCA, 2002*)

($\bar{X}$ = 116.7, σ = 15.17, C.V. = 13%)

57. For two firms A and B belonging to the same industry, the following details are available.

	Firm A	*Firm B*
No. of Employees	100	200
Average monthly wage	Rs. 2400	Rs. 1800
S.D.	Rs. 60	Rs. 80

(*i*) Which firm pays out larger amount as wages ?

(*ii*) Which firm shows greater variability in the distribution of wages ?

(*iii*) Find average monthly wage and standard deviation of all the employees in both the firms.

(*MBA, D.U. 2006*)

*****

6

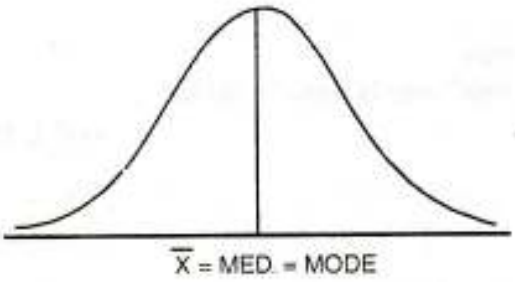
Skewness, Moments and Kurtosis

INTRODUCTION

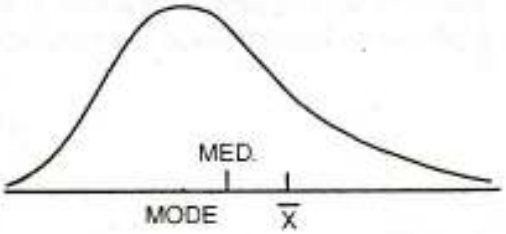
The measures of central tendency and variation discussed in previous chapters do not reveal the entire story about a frequency distribution. Two distributions may have the same mean and standard deviation but may differ in their shape of the distribution. Further description of their characteristics is necessary that is provided by measures of skewness and kurtosis.

The term 'skewness' refers to lack of symmetry or departure from symmetry, *e.g.*, when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution. The measures of skewness indicate the difference between the manner in which the observations are distributed in a particular distribution compared with a symmetrical (or normal) distribution. The concept of skewness gains importance from the fact that statistical theory is often based upon the assumption of the normal distribution. A measure of skewness is, therefore, necessary in order to guard against the consequence of this assumption.

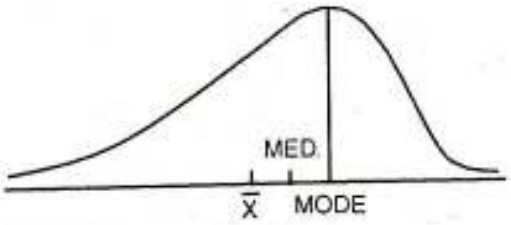
In a symmetrical distribution the values of mean, median and mode are alike. In a skewed distribution these values differ. If the value of mean is greater than the mode, skewness is said to be positive. On the other hand, if the value of mode is greater than mean, skewness is said to be negative. The following diagrams would clarify the meaning of skewness.



(*a*) Symmetrical Distribution



(*b*) Positively Skewed Distribution



(*c*) Negatively Skewed Distribution

It is clear from the (*a*), (*b*) and (*c*) diagrams that

1. In a symmetrical distribution, the values of mean, median and mode are alike.
2. In a positively skewed distribution, mean is greater than the mode and the median lies* somewhere in between mean and mode. A positively skewed distribution contains some values that are much larger than the majority of other observations.
3. In a negatively skewed distribution, mode is greater than the mean and the median lies in between mean and mode. The mean is pulled towards the low-valued item (that is, to the left). A negatively skewed distribution contains some values that are much smaller than the majority of observations.

In moderately asymmetrical distributions, the interval between the mean and the median is approximately one-third of the interval between the mean and the mode. It is this relationship that provides a means of measuring the degree of skewness.

Difference between Variation and Skewness

The following two points of difference between variation and skewness should be carefully noted :

1. Variation tells us about the amount of the variation. Skewness tells us about the direction of variation.
2. In business and economic series, measures of variation have greater practical applications than measures of skewness.

Measures of Skewness

Measures of skewness can be both absolute as well as relative. Since in a symmetrical distribution mean, median and mode are identical, the more the mean moves away from the mode, the larger the asymmetry or skewness. The distance between the mean and the mode is Karl Pearson's basis for measuring skewness. However, a measure of absolute skewness cannot be used for purposes of comparison because the same amount of skewness has different meanings in distribution with small variation and in distributions with large variation. In order to make valid comparison between the skewness in two or more distributions, we have to eliminate the distributing influence of variation. Such elimination is accomplished by dividing the absolute skewness by standard deviation. The following are two important methods of measuring relative skewness :

1. *Karl Pearson's Coefficient of Skewness.* The method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is as follows :

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Sk_p = Pearsonian (or Karl Pearson's) coefficient of skewness.

The Pearsonian coefficient of skewness is based on the same relationship as the formula for the empirical mode. The direction of skewness is determined by observing whether the mean is greater than the mode (positive skewness) or less than the mode (negative skewness). The extent of departure from symmetry is ascertained by observing the extent to which the mean is pulled away from the mode. The extent of departure is expressed in standard units in order to obtain a measure that is *independent* of the unit of measurement.

*The distance between the mode and the median is twice the distance between the median and the mean.

As the departure from symmetry becomes substantial, the relationship on which the Pearsonian coefficient formula is based breaks down and the Pearsonian coefficient no longer provides reliable results.

The value of this coefficient would be zero in a symmetrical distribution. If mean is greater than mode, coefficient of skewness would be positive, otherwise negative. In practice, the value of this coefficient usually lies between ±1 for moderately skewed distribution.

If the mode is ill-defined, the above formula has to be modified. In such a case the following approximate formula is used :

$$Sk_P = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

2. *Bowley's Coefficient of Skewness.* This method is based on quartiles. The formula for calculating coefficient of skewness is :

$$Sk_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2\text{Med.}}{Q_3 - Q_1}$$

The value of this coefficient will be zero if it is a symmetrical distribution. If the value is greater than zero, it is positively skewed and if the value is less than zero, it is negatively skewed distribution.

Sk_B= Bowley's coefficient varies between ± 1 for moderately skewed distribution.

This method is particularly useful in case of open-end distributions and where extreme values are present. Also when positional measures are called for, skewness should be measured by the Bowley's method.

3. *Kelly's Coefficient of Skewness.* Another measure of skewness devised by Kelly is based on percentiles and deciles.

The formula for calculating coefficient of skewness is given below :

$$Sk_K = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \quad \text{(based on percentiles)}$$

$$Sk_K = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1} \quad \text{(based on deciles)}$$

Sk_K = Kelly's coefficient of skewness.

It is clear from this formula that to calculate coefficient of skewness we have to determine the value of 10th, 50th and 90th percentiles. However, this method is not very popular in practice.

It should be noted that three different formulae of calculating skewness are based on different assumptions and hence the answer obtained from the same question by different method may differ.

It may be pointed out that measures of coefficient of skewness are used mainly for making comparison between two or more distributions. As a description of one distribution alone, the interpretation of a measure of skewness is vague as 'slight skewness', 'marked skewness', or 'moderate skewness'.

Illustration 1. The following data relate to the profits of 1,000 companies :

Profits (Rs. lakhs)	*No. of companies*	*Profits (Rs. lakhs)*	*No. of companies*
100–120	17	180–200	327
120–140	53	200–220	208
140–160	199	220–240	2
160–180	194		

Calculate the coefficient of skewness and comment on its value. *(MBA, M.D. Univ., 2001)*

Solution. CALCULATION OF COEFFICIENT OF SKEWNESS

Profits (Rs. lakhs)	*m.p. X*	*f*	*(X–170)/20 d*	*fd*	*fd²*
100–120	110	17	–3	–51	153
120–140	130	53	–2	–106	212
140–160	150	199	–1	–199	199
160–180	**170**	194	0	0	0
180–200	190	327	+1	+327	327
200–220	210	208	+2	+416	832
220–240	230	2	+3	+6	18
		$N = 1{,}000$		$\Sigma fd = 393$	$\Sigma fd^2 = 1{,}741$

$$Sk_P = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Calculation of Mean : $\bar{X} = A + \frac{\Sigma fd}{N} \times i = 170 + \frac{393}{1000} \times 20 = 170 + 7.86 = 177.86$

Calculation of Mode : By inspection mode lies in the class 180–200.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 180 + \frac{133}{133+119} \times 20 = 180 + 10.56 = 190.56$$

Calculation of Standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{1741}{1000} - \left(\frac{393}{1000}\right)^2} \times 20$$

$$= \sqrt{1.74 - 0.15} \times 20 = 1.26 \times 20 = 25.2$$

$$Sk_P = \frac{177.86 - 190.56}{25.2} = -0.504$$

The mode is greater than the mean by an amount equal to about 50.4 per cent of the value of standard deviation. It is a case of moderate negatively skewed distribution.

Illustration 2. The following table gives the distribution of daily wages of 500 skilled workers in a factory :

Daily wages (Rs.)	*No. of workers*
Below 200	10
200–250	25
250–300	145
300–350	220
350–400	70
400 and above	30

(*i*) Obtain the limits of daily wages of central 50 per cent of the observed workers.
(*ii*) Calculate Bowley's Coefficient of Skewness. *(MBA, Delhi Univ., 2002)*

Solution. CALCULATION OF LIMITS OF CENTRAL 50% OF WORKERS AND BOWLEY'S COEFFICIENT

Daily wages (Rs.)	*f*	*c.f.*
Below 200	10	10
200–250	25	35
250–300	145	180
300–350	220	400
350–400	70	470
400 and above	30	500

For obtaining the limits of central 50% of the workers, calculate Q_1 and Q_3

$$Q_1 = \text{Size of } \frac{N}{4}\text{th observation} = \frac{500}{4} = 125\text{th observation}$$

Q_1 lies in the class 250–300.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 250 + \frac{125-35}{145} \times 50 = 250 + 31.03 = 281.03$$

$$Q_3 = \text{Size of } \frac{3N}{4}\text{th observation} = \frac{3\times 500}{4} = 375\text{th observation}$$

Q_3 lies in the class 300–350.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i = 300 + \frac{375-180}{220} \times 50 = 300 + 44.32 = 344.32$$

Hence the daily wages of central 50% of workers lies between Rs. 281.03 and Rs. 344.32.

(*ii*) Bowley's Coefficient of Skewness

$$Sk_B = \frac{Q_3 + Q_1 - 2\,\text{Med.}}{Q_3 - Q_1}$$

$$\text{Med.} = \text{Size of } \frac{N}{2}\text{th observation} = \frac{500}{2} = 250\text{th observation}$$

Median lies in the class 300–350.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i = 300 + \frac{250-180}{220} \times 50 = 300 + 15.9 = 315.9$$

$$Sk_B = \frac{344.32 + 281.03 - 2(315.9)}{344.32 - 281.03} = \frac{-6.45}{63.29} = -0.102$$

The negative coefficient –0.102 indicates that the distance between Q_3 and Q_1 is smaller than that between Q_2 and Q_1. Thus the distribution is skewed to the left or at smaller values on the X-scale.

Illustration 3. You are given the position in a factory before and after the settlement of an industrial dispute. Comment on the gains or losses from the point of view of workers and that of management :

	Before	*After*
No. of workers	3,000	2,950
Mean wage (Rs.)	2,220	2,280
Median wage (Rs.)	2,250	2,225
Standard deviation (Rs.)	300	260

Solution. The following comments can be made on the basis of information given :

(*i*) By comparing the total wage bill, we can comment on the increase or decrease in the level of wages.

	Before	*After*
Total wage bill :	3,000 × 2220 = Rs. 66,60,000	2950 × 2280 = Rs. 67,26,000

Hence the total wage bill has gone up after the settlement of dispute even though the number of workers has decreased from 3,000 to 2,950. This means that average wage is now better. This is definitely a gain to the workers. Conversely, we cannot say that increased wage bill is a loss to management because if it results in greater efficiency of workers and, therefore, higher productivity, it would be a gain to management also.

(*ii*) Median wage before settlement of the dispute was Rs. 2,250 and after settlement is Rs. 2,225. This means that formerly 50% of workers used to get wages above Rs. 2,250 and now after the settlement of dispute they get only Rs. 2,225.

(*iii*) By comparing the coefficient of variation, we can comment on the distribution pattern of wages.

	Before	*After*
Coefficient of variation :	$\frac{300}{2220} \times 100 = 13.51$	$\frac{260}{2280} \times 100 = 11.40$

Since the coefficient of variation has decreased from 13.51 to 11.40, there is sufficient evidence to conclude that wages are more uniformly distributed after the settlement of dispute, or, in other words, there is lesser inequality in the distribution of wages after the dispute is settled.

(*iv*) By comparing skewness we can comment on the nature of the distribution.

	Before	*After*
Coefficient of skewness :	$\frac{3(2220-2250)}{300} = -0.3$	$\frac{3(2280-2225)}{260} = +0.635$

The distribution was negatively skewed before the settlement and is positively skewed after the settlement.

MOMENTS

Moments are popularly used to describe the characteristic of a distribution. They represent a convenient and unifying method for summarizing many of the most commonly used descriptive statistical measures such as central tendency, variation, skewness and kurtosis. The Greek letter μ (read as mu) is generally used to denote the moments.

For Ungrouped Data

The *r*th moment of a variable X about the arithmetic mean $\bar{X}$ is given by :

$$\mu_r = \frac{1}{N} \Sigma \left(X - \bar{X}\right)^r \quad ...(i)$$

The *r*th moment of a variable X about any arbitrary point A is given by :

$$\mu'_r = \frac{1}{N} \Sigma (X - A)^r \quad ...(ii)$$

For Grouped Data

$$\mu_r = \frac{1}{N} \Sigma f (X - \bar{X})^r \quad ...(iii)$$

and

$$\mu'_r = \frac{1}{N} \Sigma f (X - A)^r \quad ...(iv)$$

For different values of r, we shall get different moments. Thus if we put $r = 1$, we will get first moment, if we put $r = 2$, we will get second moment, and so on.

Moments about Mean*

For ungrouped data :

$$\mu_1 = \frac{\Sigma (X - \bar{X})}{N}; \qquad \mu_2 = \frac{\Sigma (X - \bar{X})^2}{N}$$

$$\mu_3 = \frac{\Sigma (X - \bar{X})^3}{N}; \qquad \mu_4 = \frac{\Sigma (X - \bar{X})^4}{N}$$

For grouped data :

$$\mu_1 = \frac{\Sigma f (X - \bar{X})}{N}; \qquad \mu_2 = \frac{\Sigma f (X - \bar{X})^2}{N}$$

$$\mu_3 = \frac{\Sigma f (X - \bar{X})^3}{N}; \qquad \mu_4 = \frac{\Sigma f (X - \bar{X})^4}{N}$$

We can extend the moments to higher power in the similar way. But in practice the first four moments suffice.

The first moment about the origin tells us about the mean, the second moment about variance, the third moment about skewness and the fourth moment about the kurtosis.

*Moments about mean are also called central moments.

Moments about Arbitrary Point

When actual mean is in fraction, moments are first calculated about an arbitrary origin and then converted to moments about the actual mean. When deviations are taken from arbitrary point, the formulae are :

$$\mu'_1 = \frac{\Sigma (X - A)}{N} \qquad \mu'_2 = \frac{\Sigma (X - A)^2}{N}$$

$$\mu'_3 = \frac{\Sigma (X - A)^3}{N} \qquad \mu'_4 = \frac{\Sigma (X - A)^4}{N}$$

μ'_1, μ'_2, etc., denote first, second moment, etc., about an arbitrary point 'A'.

In a frequency distribution, to simplify calculations we can take a common factor but in that case the various moments have to be multiplied by i, i^2, i^3 and i^4 respectively. Thus, taking $d = \frac{X - A}{i}$ or $(X - A) = id$, we get

$$\mu'_1 = \frac{\Sigma f (X - A)}{N} \quad \text{or} \quad \frac{\Sigma fd}{N} \times i$$

$$\mu'_2 = \frac{\Sigma f (X - A)^2}{N} \quad \text{or} \quad \frac{\Sigma fd^2}{N} \times i^2$$

$$\mu'_3 = \frac{\Sigma f (X - A)^3}{N} \quad \text{or} \quad \frac{\Sigma fd^3}{N} \times i^3$$

$$\mu'_4 = \frac{\Sigma f (X - A)^4}{N} \quad \text{or} \quad \frac{\Sigma fd^4}{N} \times i^4$$

However, when we calculate the values of β_1 and β_2, the answer will remain the same whether we have multiplied the moments by common factor or not.

Finding Central Moments from Moments about Arbitrary Point

With the help of following relationships, moments about an arbitrary point can be converted to moments about mean :

$$\mu_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1$$

Two important constants calculated from μ_2, μ_3 and μ_4 are :

(*i*) β_1 (read as beta one) and (*ii*) β_2 (read as beta two)

(*i*) β_1 is defined as : $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

β_1 is used as a measure of skewness. In a symmetrical distribution β_1 shall be zero. However, the coefficient β_1 as a measure of skewness has a serious limitation. β_1 as a measure of skewness cannot tell us about the direction of skewness, *i.e.*, whether it is positive or negative. This is for the simple reason that μ_3 being the sum of the cubes of the deviation from the mean may be positive or negative but μ_3^2 is always positive. Also μ_2 being the variance is always positive. Hence $\beta_1 = \mu_3^2/\mu_2^3$ is always positive. This drawback is removed if we calculate Karl Pearson's γ_1 (pronounced as Gamma one). γ_1 is defined as the square root of β_1, *i.e.*,

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

The sign of skewness would depend upon the value of μ_3. If μ_3 is positive we will have positive skewness and if μ_3 is negative, we will have negative skewness.

It is advisable to use γ_1 as a measure of skewness.

(*ii*) β_2 measures kurtosis and is defined as : $\beta_2 = \frac{\mu_4}{\mu_2^{\ 2}}$.

Illustration 4. From the following data calculate first four moments and also find the value of γ_1 :

Monthly Profits (in lakh Rs.)	*No. of Companies*	*Monthly Profits (in lakh Rs.)*	*No. of Companies*
Less than 7.5	4	22.5–27.5	16
7.5–12.5	10	27.5–32.5	12
12.5–17.5	20	32.5–37.5	2
17.5–22.5	36		

Solution. CALCULATION OF MOMENTS

Monthly Profits (in lakh Rs.)	*m.p.* X	f	$(X-20)/5$ d	fd	fd^2	fd^3	fd^4
Less than 7.5	5	4	–3	–12	36	–108	324
7.5–12.5	10	10	–2	–20	40	–80	160
12.5–17.5	15	20	–1	–20	20	–20	20
17.5–22.5	20	36	0	0	0	0	0
22.5–27.5	25	16	+1	+16	16	+16	16
27.5–32.5	30	12	+2	+24	48	+96	192
32.5–37.5	35	2	+3	+6	18	+54	162
		$N = 100$		Σfd $= -6$	Σfd^2 $= 178$	Σfd^3 $= -42$	Σfd^4 $= 874$

Moments about arbitrary origin (20) in class-interval units :

$$\mu'_1 = \frac{\Sigma fd}{N} \times i = \frac{-6}{100} \times 5 = -0.3; \quad \mu'_2 = \frac{\Sigma fd^2}{N} \times i^2 = \frac{178}{100} \times 25 = 44.5;$$

$$\mu'_3 = \frac{\Sigma fd^3}{N} \times i^3 = \frac{-42}{100} \times 125 = -52.5; \quad \mu'_4 = \frac{\Sigma fd^4}{N} \times i^4 = \frac{874}{100} \times 625 = 5462.5$$

Moments about mean

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$
$$= 44.5 - (-0.3)^2 = 44.5 - 0.09 = 44.41$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3$$
$$= -52.5 - 3(-0.3 \times 44.5) + 2(-0.3)^3$$
$$= -52.5 + 40.05 - .054 = -12.504$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2\mu'^{\ 2}_1 - 3\mu'^{\ 4}_1$$
$$= 5462.5 - 4(-0.3 \times -52.5) + 6(44.5)(-0.3)^2 - 3(-0.3)^4$$
$$= 5462.5 - 63 + 24.03 - .0243 = 5423.5057$$

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{-12.504}{(6.6641)^3} = -\frac{12.504}{295.954} = -0.0422.$$

Illustration 5. The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Show that the mean is 7. Also find the other moments and β_1 and β_2.

Solution. We are given

$$\mu'_1 = 2,\ \mu'_2 = 20,\ \mu'_3 = 40 \text{ and } \mu'_4 = 50 \text{ and } A = 5.$$

We have to find the moments about mean.

$$\overline{X} = \mu'_1 + A = 2 + 5 = 7$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 20 - (2)^2 = 16$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'^3_1 = 40 - 3\,(2)\,(20) + 2\,(2)^3 = -64$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 = 50 - 4\,(40)\,(2) + 6\,(20)\,(2)^2 - 3\,(2)^4 = 162$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-64)^2}{(16)^3} = \frac{4096}{4096} = +1.00$$

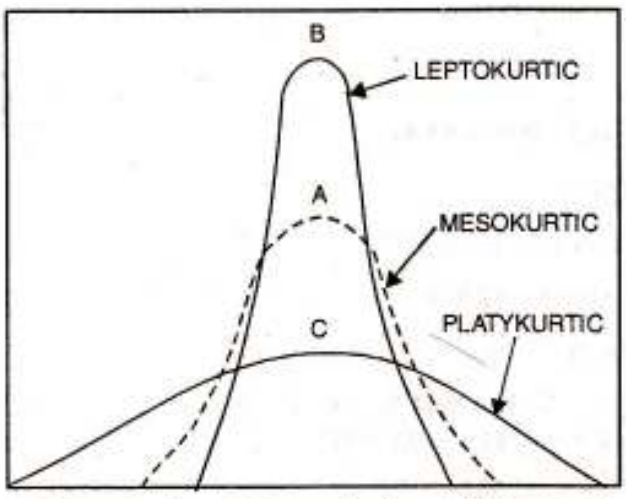
$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{162}{(16)^2} = \frac{162}{256} = +0.63$$

KURTOSIS

In describing a frequency distribution, a person can use an average to show the typical value or central tendency in the distribution, a measure of variation to show the variation of values either with certain values (such as the range and quartile deviation) or around the average of the distribution (such as the average deviation and the standard distribution) either skewed to the higher values (the right side on the X-scale) or to the lower values (the left side on the X-scale). Further, the measure of kurtosis, the fourth device in describing a frequency distribution, can be used to show the degree of concentration, either the values concentrated in the area around the mode (a peaked curve) or decentralised from the mode of both tails of the frequency curve (a flat topped curve).

Kurtosis in Greek means "*bulginess*". In statistics, kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of a normal curve. If a curve is more peaked than the normal curve, it is called 'leptokurtic' ; if it is more or flat-topped than the normal curve, it is called 'platykurtic' or flat-topped. The normal curve itself is known as 'mesokurtic'. The concept of kurtosis is rarely used in analysing business data :

The diagram below illustrates the scope of three different curves mentioned above :



(*A*) Mesokurtic. (*B*) Leptokurtic. (*C*) Platykurtic.

Measures of Kurtosis

Kurtosis is measured by β_2 or its derivative γ_2

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ and } \gamma_2 = \beta_2 - 3.$$

For a symmetrical (normal) distribution the value of $\beta_2=3$ [or $\gamma_2=0$]. If the value of β_2 is greater than 3, the curve is more peaked than the normal curve; *i.e.,* leptokurtic; when the value of β_2 is less than 3, the curve is less peaked than normal curve *i.e.,* platykurtic. It may be noted that it is easier to interpret kurtosis by calculating β_2 instead of γ_2.

Illustration 6. The first central moments of a distribution are 0, 16, –36 and 120. Comment on the skewness and kurtosis of the distribution.

Solution. We are given $\mu_1 = 0$, $\mu_2 = 16$, $\mu_3 = -36$ and $\mu_4 = 120$. For commenting on the skewness we calculate γ_1.

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{-36}{(4)^3} = \frac{-36}{64} = -0.5625 \qquad \sigma = \sqrt{\mu_2} = \sqrt{16} = 4$$

The distribution is negatively skewed (It may be noted that if we calculate β_1 its value will be $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-36)^2}{(16)^3} = +0.3164$. But this would be wrong as μ_3 is negative). For commenting on the kurtosis we calculate β_2.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{120}{(16)^2} = +0.469$$

Since the value of β_2 is less than 3, the distribution is platykurtic.

MISCELLANEOUS ILLUSTRATIONS

Illustration 7. An analysis of production rejects resulted in the following figures :

No. of rejects per operator	*No. of operators*	*No. of rejects per operator*	*No. of operators*
21–25	5	41–45	15
26–30	15	46–50	12
31–35	28	51–55	3
36–40	42		

Calculate mean, standard deviation and coefficient of skewness and comment on the results.

Solution. COMPUTATION OF COEFFICIENT OF SKEWNESS

No. of rejects per operator	*m.p.* X	*No. of operators* f	$(X-38)/5$ d	fd	fd^2
20.5–25.5	23	5	–3	–15	45
25.5–30.5	28	15	–2	–30	60
30.5–35.5	33	28	–1	–28	28
35.5–40.5	**38**	42	0	0	0
40.5–45.5	43	15	+1	+15	15
45.5–50.5	48	12	+2	+24	48
50.5–55.5	53	3	+3	+9	27
		$N = 120$		$\Sigma fd = -25$	$\Sigma fd^2 = 223$

$$\text{Mean: } \bar{X} = A + \frac{\Sigma fd}{N} \times i = 38 - \frac{25}{120} \times 5 = 38 - 1.04 = 36.96$$

$$\text{Standard deviation : } \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{223}{120} - \left(\frac{-25}{120}\right)^2} \times 5$$

$$= \sqrt{1.8583 - .0434} \times 5 = \sqrt{1.8149} \times 5 = 1.3472 \times 5 = 6.736$$

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 35.5 + \frac{14}{14+27} \times 5 = 35.5 + 1.71 = 37.21$$

$$\text{Hence Coeff. of } Sk = \frac{36.96 - 37.21}{6.736} = \frac{-.25}{6.736} = -0.037$$

The value of mean = 36.96 indicates that on the average, rejects per operator were 37 in number. The value of standard deviation = 6.736 suggests that the variation in the data from the central value is approximately 7. Coefficient of skewness = –0.037 indicates that the distribution is slightly skewed to the left and therefore, there is greater concentration of the rejects per operator at the upper values than the lower values of the distribution.

Illustration 8. Distinguish between Karl Pearson's and Bowley's coefficient of skewness. Compute an appropriate measure of skewness for the following data:

Sales (Rs. Lakhs)	*No. of Companies*	*Sales (Rs. Lakhs)*	*No. of Companies*
Below 50	12	90–100	55
50–60	30	100–110	45
60–70	65	110–120	25
70–80	78	Above 120	10
80–90	80		

Solution. Since it is an open-end distribution, therefore Bowley's method of calculating skewness should be more appropriate.

CALCULATION OF COEFFICIENT OF SKEWNESS

Sales	*f*	*c.f.*
Belew 50	12	12
50–60	30	42
60–70	65	107
70–80	78	185
80–90	80	265
90–100	55	320
100–110	45	365
110–120	25	390
Above 120	10	400

$$\text{Coeff. of } Sk = \frac{Q_3 + Q_1 - 2 \text{ Med.}}{Q_3 - Q_1}$$

Q_1 = Size of $\frac{N}{4}$th observation $= \frac{400}{4} = 100$th observation.

Q_1 lies in the class 60–70.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 60 + \frac{100 - 42}{65} \times 10 = 60 + 8.92 = 68.92$$

Q_3 = Size of $\frac{3N}{4}$th observation $= \frac{3 \times 400}{4} = 300$th observation.

Q_3 lies in the class 90–100.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i = 90 + \frac{300 - 265}{55} \times 10 = 90 + 6.36 = 96.36$$

$$\text{Med.} = \text{Size of } \frac{N}{2}\text{th observation} = \frac{400}{2} = 200\text{th observation}$$

Median lies in the class 80–90.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i = 80 + \frac{200 - 185}{80} \times 10 = 80 + 1.875 = 81.875$$

$$\text{Coeff. of } Sk = \frac{96.36 + 68.92 - 2(81.875)}{96.36 - 68.92} = \frac{165.28 - 163.75}{27.44} = 0.056.$$

Illustration 9. Find an appropriate measure of skewness from the following distribution :

Age (yrs.)	*No. of employees*	*Age (yrs.)*	*No. of employees*
Below 20	13	35–40	112
20–25	29	40–45	94
25–30	46	45–50	45
30–35	60	50 and above	21

(MBA, Bharthidasan Univ., 2007)

Solution. Since it is an open-end distribution, therefore appropriate measure of skewness would be Bowley's coefficient of skewness.

CALCULATION OF BOWLEY'S COEFFICIENT

Age (Yrs.)	*No. of employees (f)*	*c.f.*
Below 20	13	13
20–25	29	42
25–30	46	88
30–35	60	148
35–40	112	260
40–45	94	354
45–50	45	399
50 and above	21	420
	$N = 420$	

$$Sk_B = \frac{Q_3 + Q_1 - 2\,\text{Med.}}{Q_3 - Q_1}$$

Q_1 = Size of $\frac{N}{4}$th observation = $\frac{420}{4}$ = 105th observation

Q_1 lies in the class 30–35.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 30 + \frac{105 - 88}{60} \times 5 = 30 + 1.42 = 31.42$$

Q_3 = Size of $\frac{3N}{4}$th observation = $\frac{3 \times 420}{4}$ = 315th observation

Q_3 lies in the class 40–45.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i = 40 + \frac{315 - 260}{94} \times 5 = 40 + 2.93 = 42.93$$

Med. = Size of $\frac{N}{2}$th observation = $\frac{420}{2}$ = 210th observation

Median lies in the class 35–40.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i = 35 + \frac{210 - 148}{112} \times 5 = 35 + 2.77 = 37.77$$

$$Sk_B = \frac{42.93 + 31.42 - (2 \times 37.77)}{42.93 - 31.42} = \frac{-1.19}{11.51} = -0.103$$

Illustration 10. (*a*) The sum of 50 observations is 500, its sum of squares is 6,000 and median 12. Find the coefficient of variation and coefficient of skewness.

Solution. $N = 50$, $\Sigma X = 500$, $\Sigma X^2 = 6{,}000$, Med. = 12

$$\bar{X} = \frac{\Sigma X}{N} = \frac{500}{50} = 10 \text{ ; and } \sigma = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2} = \sqrt{\frac{6{,}000}{50} - (10)^2} = 4.47$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{4.47}{10} \times 100 = 44.7 \text{ per cent}$$

Mode = 3 Median – 2 Mean = 3 × 12 – 2 × 10 = 16

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{10 - 16}{4.47} = -1.34.$$

(*b*) For a moderately skewed distribution, the arithmetic mean is 100 and coefficient of variation is 35, and Pearson's coefficient of skewness is 0.2. Find the mode and the median.

Solution. $\bar{X} = 100$, C.V. = 35 $Sk_P = 0.2$.

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100$$

$$35 = \frac{\sigma}{100} \times 100 \text{ or } \sigma = 35$$

$$Sk_P = \frac{\bar{X} - \text{Mode}}{\sigma} \text{ or } 0.2 = \frac{100 - \text{Mode}}{35}$$

$$7 = 100 - \text{Mode or Mode} = 93$$

$$\text{Mode} = 3 \text{ Med.} - 2 \text{ Mean}$$

$$93 = 3 \text{ Med.} - 2 \times 100 \text{ or } 3 \text{ Med.} - 200 = 93$$

$$3 \text{ Med.} = 293 \quad \therefore \text{ Med.} = 97.7$$

Hence Mode = 93 and Median = 97.7

Illustration 11. From the following data of age of employees, calculate coefficient of skewness and comment on the result :

Age below (yrs.) :	25	30	35	40	45	50	55
No. of employees :	8	20	40	65	80	92	100

Solution. This is a cumulative frequency distribution. First we will convert it to a simple frequency distribution and then calculate coefficient of skewness.

CALCULATION OF COEFFICIENT OF SKEWNESS

Age (*Yrs.*)	*m.p.* X	f	$(X-37.5)/5$ d	fd	fd^2
20–25	22.5	8	−3	−24	72
25–30	27.5	12	−2	−24	48
30–35	32.5	20	−1	−20	20
35–40	37.5	25	0	0	0
40–45	42.5	15	+1	+15	15
45–50	47.5	12	+2	+24	48
50–55	52.5	8	+3	+24	72
		$N = 100$		$\Sigma fd = -5$	$\Sigma fd^2 = 275$

$$Sk_P = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Mean :
$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 37.5 - \frac{5}{100} \times 5 = 37.25$$

Mode : Mode lies in the class 35 – 40.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 35 + \frac{5}{5+10} \times 5 = 36.67$$

$$S.D. : \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{275}{100} - \left(\frac{-5}{100}\right)^2} \times 5$$

$$= \sqrt{2.75 - .0025} \times 5 = 1.658 \times 5 = 8.29$$

$$Sk_P = \frac{37.25 - 36.67}{8.29} = \frac{0.58}{8.29} = 0.07.$$

This value of skewness indicates that the distribution has hardly any skewness.

Illustration 12. You are given the following frequency distribution of the daily earnings of employees in a company :

Earnings (*in Rs.*)	*Number of workers*	*Earnings* (*in Rs.*)	*Number of workers*
50–70	4	130–150	6
70–90	8	150–170	7
90–110	12	170–190	3
110–130	20		

Calculate the first four moments about the point 120. Convert the result into moments about the mean. Compute the value of γ_1 and γ_2 and comment on the result. *(MBA, Delhi Univ., 2002)*

Solution. Moment about some arbitrary point is given by

$$\mu_r' = \frac{1}{N} \Sigma f(X-A)^r$$

Here $A = 120$ and X are the mid-points. To get the first four moments, put $r = 1, 2, 3$ and 4 in the above formula.

COMPUTATION OF FIRST FOUR MOMENTS

Earnings (Rs.)	*m.p.* X	f	$(X-120)/20$ d	fd	fd^2	fd^3	fd^4
50–70	60	4	−3	−12	36	−108	324
70–90	80	8	−2	−16	32	−64	128
90–110	100	12	−1	−12	12	−12	12
110–130	**120**	20	0	0	0	0	0
130–150	140	6	+1	+6	6	+6	6
150–170	160	7	+2	+14	28	+56	112
170–190	180	3	+3	+9	27	+81	243
		$N = 60$		$\Sigma fd = -11$	$\Sigma fd^2 = 141$	$\Sigma fd^3 = -41$	$\Sigma fd^4 = 825$

Moments about the arbitrary point = 120

$$\mu'_1 = \frac{\Sigma fd}{N} \times i = \frac{-11}{60} \times 20 = -3.6667$$

$$\mu'_2 = \frac{\Sigma fd^2}{N} \times i^2 = \frac{141}{60} \times (20)^2 = 940$$

$$\mu'_3 = \frac{\Sigma fd^3}{N} \times i^3 = \frac{-41}{60} \times (20)^3 = -5466.6667$$

$$\mu'_4 = \frac{\Sigma fd^4}{N} \times i^4 = \frac{825}{60} \times (20)^4 = 22,00,000$$

Moments about mean :

$\mu_1 = 0$ (since the sum of the deviations from the means is zero.)

$\mu_2 = \mu'_2 - \mu'^2_1 = 940 - (-3.6667)^2 = 926.5553$ or $\sigma = \sqrt{\mu_2} = 30.4394$

$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'^3_1$

$= -5466.6667 - 3(940)(-3.6667) + 2(-3.6667)^3$

$= -5466.6667 + 10340.094 - 98.5953 = 4774.832$

$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2\mu'^2_1 - 2\mu'^4_1$

$= 2200000 - 4(-3.6667)(-5466.6667) + 6(940)(-3.6667)^2 - 3(-3.6667)^4$

$= 2200000 - 80178.507 + 75828.045 - 542.2789 = 2195107.3$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3} = \frac{4774.83}{(30.4394)^3} = 0.1693;$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{2195107.3}{(926.56)^2} = 2.56, \quad \gamma_2 = \beta_2 - 3 = -0.44$$

The value of γ_1 indicates that the distribution is slightly skewed to the right, *i.e.*, it is not perfectly symmetrical. Since the value of γ_2 is less than zero, therefore, the distribution is platykurtic.

Illustration 13. (*a*) The first three moments of a distribution about the value 1 are 2, 25 and 80. Find its mean, standard deviation and the moment-measure of skewness.

Solution. $\mu'_1 = 2, \ \mu'_2 = 25, \ \mu'_3 = 80, \ A = 1$

Mean : $\bar{X} = \mu'_1 + A = 2 + 1 = 3$

Standard deviation : $\mu_2 = \mu'_2 - \mu'^2_1 = 25 - (2)^2 = 21$

$\sigma = \sqrt{\mu_2} = \sqrt{21} = 4.583$

$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3 = 80 - 3 \times 2 \times 25 + 2(2)^3$ or $\mu_3 = 80 - 150 + 16 = -54$

Moment-measure of skewness : $\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3} = \frac{(-54)}{(4.583)^3} = \frac{-54}{96.26} = -0.561$

(*b*) The first and second moment of a distribution about the value 5 of the variable are 2 and 20. Find the mean and standard deviation.

Solution.

$$\mu'_1 = 2,\ \mu'_2 = 20,\ A = 5$$

$$\bar{X} = \mu'_1 + A = 2 + 5 = 7$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 20 - (2)^2 = 16$$

$$\sigma = \sqrt{\mu_2} = \sqrt{16} = 4.$$

Illustration 14. Find the second, third and the fourth central moments of the frequency distribution given below. Hence find (*i*) a measure of skewness, and (*ii*) a measure of kurtosis.

Class Limits	*Frequency*	*Class Limits*	*Frequency*
110.0–114.9	5	130.0–134.9	10
115.0–119.9	15	135.0–139.9	10
120.0–124.9	20	140.0–144.9	5
125.0–129.9	35		

Solution.

CALCULATION OF MOMENTS

Class Limits	*m.p.* X	f	$(X-127.45)/5$ d	fd	fd^2	fd^3	fd^4
110.0–114.9	112.45	5	–3	–15	45	–135	405
115.0–119.9	117.45	15	–2	–30	60	–120	240
120.0–124.9	122.45	20	–1	–20	20	–20	20
125.0–129.9	127.45	35	0	0	0	0	0
130.0–134.9	132.45	10	+1	+10	10	+10	10
135.0–139.9	137.45	10	+2	+20	40	+80	160
140.0–144.9	142.45	5	+3	+15	45	+135	405
		$N = 100$		$\Sigma fd = -20$	$\Sigma fd^2 = 220$	$\Sigma fd^3 = -50$	$\Sigma fd^4 = 1{,}240$

$$\mu'_1 = \frac{\Sigma fd}{N} \times i = \frac{-20}{100} \times 5 = -1; \qquad \mu'_2 = \frac{\Sigma fd^2}{N} \times i^2 = \frac{220}{100} \times 25 = 55$$

$$\mu'_3 = \frac{\Sigma fd^3}{N} \times i^3 = \frac{-50}{100} \times 125 = -62.5; \qquad \mu'_4 = \frac{\Sigma fd^4}{N} \times i^4 = \frac{1240}{100} \times 625 = 7750$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 55 - (-1)^2 = 55 - 1 = 54 \text{ or } \sigma = \sqrt{\mu_2} = 7.348$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'^3_1 = -62.5 - 3(-1)(55) + 2(-1)^3$$

$$= -62.5 + 165 - 2 = 100.5$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1$$

$$= 7750 - 4(-1)(-62.5) + 6(55)(-1)^2 - 3(-1)^4$$

$$= 7750 - 250 + 330 - 3 = 7827$$

Measure of skewness : $\gamma_1 = \dfrac{\mu_3}{\sigma^3} = \dfrac{100.5}{(7.348)^3} = \dfrac{100.5}{396.74} = +0.253$

Measure of kurtosis : $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{7827}{(54)^2} = 2.684$

Since the value of β_2 is less than 3, the curve is platykurtic.

Illustration 15. Calculate coefficient of variation and Karl Pearson's coefficient of skewness from the data given below :

Marks	*No. of students*
Less than 30	18
" " 40	40
" " 60	70
" " 80	90
" " 100	100

(MBA, Kumaun Univ., 2002)

Solution.

CALCULATION OF COEFFICIENT OF VARIATION AND COEFFICIENT OF SKEWNESS

Marks	*m.p.* X	f	$(X-50)/20$ d	fd	fd^2
0 – 20	10	18	–2	–36	72
20 – 40	30	22	–1	–22	22
40 – 60	**50**	30	0	0	0
60 – 80	70	20	+1	+20	20
80 – 100	90	10	+2	+20	40
		$N=100$		$\Sigma fd = -18$	$\Sigma fd^2 = 154$

$$\overline{X} = A + \frac{\Sigma fd}{N} \times i = 50 - \frac{18}{100} \times 20 = 50 - 3.6 = 46.4$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{154}{100} - \left(\frac{-18}{100}\right)^2} \times 20$$

$$= \sqrt{1.54 - 0.0324} \times 20 = 1.228 \times 20 = 24.56$$

$$\text{C.V.} = \frac{\sigma}{\overline{X}} \times 100 = \frac{24.56}{46.4} \times 100 = 52.93$$

By inspection mode lies in the class 40–60.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 40 + \frac{8}{8+10} \times 20 = 40 + 8.89 = 48.89$$

$$\text{Coeff. of Sk} = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{46.4 - 48.89}{24.56} = \frac{-2.49}{24.56} = -0.101.$$

Therefore, it is a case of low degree of negatively skewed distribution.

Illustration 16. Calculate Bowley's coefficient of skewness from the following data :

Sales (*Rs. Lakhs*)	*No. of Companies*
Below 50	8
" 60	20
" 70	40
" 80	65
" 90	80

Solution. *(MBA, Osmania Univ.; MBA, Delhi Univ., 2006)*

CALCULATION OF BOWLEY'S COEFFICIENT OF SKEWNESS

Sales (*Rs.Lakhs*)	*No. of Companies* f	*c.f.*
40 – 50	8	8
50 – 60	12	20
60 – 70	20	40
70 – 80	25	65
80 – 90	15	80

$$\text{Bowley's Coeff. of Sk} = \frac{Q_3 + Q_1 - 2\,\text{Med.}}{Q_3 - Q_1}$$

$$Q_1 = \text{Size of } \frac{N}{4}\text{th observation} = \frac{80}{4} = 20\text{th observation}$$

Q_1 lies in the class 50–60.

$$Q_1 = L + \frac{N/4 - p.c.f.}{f} \times i = 50 + \frac{20-8}{12} \times 10 = 50 + 10 = 60.$$

$$Q_3 = \text{Size of } \frac{3N}{4} \text{th observation} = \frac{3 \times 80}{4} = 60\text{th observation.}$$

Q_3 lies in the class 70–80.

$$Q_3 = L + \frac{3N/4 - p.c.f.}{f} \times i$$

$$= 70 + \frac{60-40}{25} \times 10 = 70 + 8 = 78$$

$$\text{Med.} = \text{Size of } \frac{N}{2} \text{th observation} = \frac{80}{2} = 40\text{th observation}$$

Median lies in the class 60–70.

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i$$

$$= 60 + \frac{40-20}{20} \times 10 = 60 + 10 = 70$$

$$\text{Coeff. of Sk} = \frac{78 + 60 - 2(70)}{78 - 60} = \frac{78 + 60 - 140}{18} = -0.111.$$

Therefore, it is a case of less negatively skewed distribution.

Illustration 17. The following table gives the length of life (in hours) of 400 T.V. picture tubes :

Length of life (in hours)	*No. of picture tubes*	*Length of life (in hours)*	*No. of picture tubes*
4000–4199	12	5000–5199	55
4200–4399	30	5200–5399	36
4400–4599	65	5400–5599	25
4600–4799	78	5600–5799	9
4800–4999	90		

Compute mean, standard deviation and coefficient of skewness. Comment on the values obtained. *(MBA, Delhi Univ.)*

Solution. CALCULATION OF MEAN, STANDARD DEVIATION AND COEFFICIENT OF SKEWNESS

Length of life (in hours)	f	*m.p.* X	$(x - 4899.5)/200$ d	fd	fd^2
4000–4199	12	4099.5	–4	–48	192
4200–4399	30	4299.5	–3	–90	270
4400–4599	65	4499.5	–2	–130	260
4600–4799	78	4699.5	–1	–78	78
4800–4999	90	**4899.5**	0	0	0
5000–5199	55	5099.5	+1	+55	55
5200–5399	36	5299.5	+2	+72	144
5400–5599	25	5499.5	+3	+75	225
5600–5799	9	5699.5	+4	+36	144
	$N = 400$			$\Sigma fd = -108$	$\Sigma fd^2 = 1368$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 4899.5 - \frac{108}{400} \times 200 = 4899.5 - 54 = 4845.5$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{1368}{400} - \left(\frac{-108}{400}\right)^2} \times 200$$

$$= \sqrt{3.42 - .0729} \times 200 = 1.8295 \times 200 = 365.9$$

$$\text{Coeff. of Sk} = \frac{\bar{X} - \text{Mode}}{\sigma}$$

Mode lies in the class 4800–4999. But the real limit of this class is 4799.5 – 4999.5.

$$Mode = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 4799.5 + \frac{12}{12+35} \times 200 = 4799.5 + 51.06 = 4850.56$$

$$\text{Coeff. of Sk} = \frac{4845.5 - 4850.56}{365.9} = \frac{-5.06}{365.9} = -0.014.$$

It is a case of very very low degree of negative skewness.

Illustration 18. You are given the following data pertaining to kilowatt hours of electricity consumed by 100 persons in Delhi :

Consumption (in K-Watt hours) :	0–10	10–20	20–30	30–40	40–50
No. of users :	6	25	36	20	13

Calculate (*i*) arithmetic mean, (*ii*) standard deviation and (*iii*) coefficient of skewness.

Solution. CALCULATION OF COEFFICIENT OF SKEWNESS

Consumption (kw. hours)	*Mid-point X*	*f*	*(X–25)/10 d*	*fd*	fd^2
0–10	5	6	–2	–12	24
10–20	15	25	–1	–25	25
20–30	**25**	36	0	0	0
30–40	35	20	+1	+20	20
40–50	45	13	+2	+26	52
		$N = 100$		$\Sigma fd = 9$	$\Sigma fd^2 = 121$

Calculation of Mean : $\bar{X} = A + \frac{\Sigma fd}{N} \times i = 25 + \frac{9}{100} \times 10 = 25.9$

Calculation of Mode. Since the highest frequency is 36, mode lies in the class 20–30.

$$Mode = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 20 + \frac{11}{11+16} \times 10 = 20 + 4.07 = 24.07$$

Calculation of S.D. :

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{121}{100} - \left(\frac{9}{100}\right)^2} \times 10$$

$$= \sqrt{1.21 - .0081} \times 10 = 10.963$$

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{25.9 - 24.07}{10.963} = 0.167.$$

Illustration 19. Calculate Karl Pearson's coefficient of skewness from the following data :

Class	*Frequency*	*Class*	*Frequency*
70–80	5	30–40	35
60–70	6	20–30	30
50–60	11	10–20	22
40–50	21	0–10	11

Solution. Arrange the class/groups and the corresponding frequencies in the ascending order.

CALCULATION OF KARL PEARSON'S COEFFICIENT OF SKEWNESS

Class	*Mid-point X*	*f*	*(X–35)/10 d*	*fd*	fd^2
0–10	5	11	–3	–33	99
10–20	15	22	–2	–44	88
20–30	25	30	–1	–30	30
30–40	**35**	35	0	0	0
40–50	45	21	+1	+21	21
50–60	55	11	+2	+22	44
60–70	65	6	+3	+18	54
70–80	75	5	+4	+20	80
		$N = 141$		$\Sigma fd = -26$	$\Sigma fd^2 = 416$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 35 - \frac{26}{141} \times 10 = 35 - 1.844 = 33.156$$

Mode lies in the class 30–40.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 30 + \frac{5}{5+14} \times 10 = 30 + 2.63 = 32.63$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{416}{141} - \left(\frac{-26}{141}\right)^2} \times 10$$

$$= \sqrt{2.95 - .034} \times 10 = 1.708 \times 10 = 17.08$$

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{33.156 - 32.63}{17.08} = \frac{0.526}{17.08} = 0.031$$

It is a very very low degree of positive skewness.

Illustration 20. The following table gives the length of life (in hours) of 400 T.V. picture tubes:

Length of life (in hours)	*No. of picture tubes*	*Length of life (in hours)*	*No. of picture tubes*
4000–4200	22	4800–5000	80
4200–4400	38	5000–5200	70
4400–4600	65	5200–5400	50
4600–4800	75		

Compute arithmetic mean, mode, standard deviation and coefficient of skewness.

Solution. CALCULATION OF $\bar{X}$, MODE, σ AND COEFFICIENT OF SKEWNESS

Length of life (in hours)	X	f	$(X-4700)/200$ d	fd	fd^2
4000–4200	4100	22	−3	−66	198
4200–4400	4300	38	−2	−76	152
4400–4600	4500	65	−1	−65	65
4600–4800	4700	75	0	0	0
4800–5000	4900	80	+1	+80	80
5000–5200	5100	70	+2	+140	2s80
5200–5400	5300	50	+3	+150	450
		N = 400		Σfd = 163	Σfd^2 = 1225

$$\textit{Mean}: \bar{X} = A + \frac{\Sigma fd}{N} \times i = 4700 + \frac{163}{400} \times 200 = 4700 + 81.5 = 4781.5$$

Mode : Mode lies in the class 4800–5000.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 4800 + \frac{5}{5+10} \times 200 = 4800 + 66.67 = 4866.67$$

$$\text{S.D.} : \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{1225}{400} - \left(\frac{163}{400}\right)^2} \times 200$$

$$= \sqrt{3.0625 - .166} \times 200 = 1.702 \times 200 = 340.4$$

$$\text{Coeff. of Sk} = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{4781.5 - 4866.67}{340.4} = \frac{-85.17}{340.4} = -0.25.$$

Illustration 21. Calculate Karl Pearson's coefficient of skewness from the following data:

Marks	*No. of students*	*Marks*	*No. of students*
above 0	150	above 50	70
" 10	140	" 60	30
" 20	100	" 70	14
" 30	10	" 80	0
" 40	75		

(MBA, M.D. Univ., 1998)

Solution. This is a cumulative frequency distribution. First convert it to a simple frequency distribution and then calculate coefficient of skewness.

CALCULATION OF KARL PEARSON'S COEFFICIENT OF SKEWNESS

Marks	*m.p.* X	f	$(X-35)/10$ d	fd	fd^2	*c.f.*
0–10	5	10	−3	−30	90	10
10–20	15	40	−2	−80	160	50
20–30	25	20	−1	−20	20	70
30–40	**35**	5	0	0	0	75
40–50	45	5	+1	+5	5	80
50–60	55	40	+2	+80	160	120
60–70	65	16	+3	+48	144	136
70–80	75	14	+4	+56	224	150
			$N=150$	$\Sigma fd = 59$	$\Sigma fd^2 = 803$	

Since the maximum frequency 40 has been repeated twice, it is a bimodal distribution and hence we will use the formula.

$$\text{Coeff. of Sk} = \frac{3(\bar{X} - \text{Med.})}{\sigma}$$

$$\textit{Mean}: \bar{X} = A + \frac{\Sigma fd}{N} \times i = 35 + \frac{59}{150} \times 10 = 35 + 3.93 = 38.93$$

$$\textit{Median}: \text{Med.} = \text{Size of } \frac{N}{2}\text{th observation} = \frac{150}{2} = 75\text{th observation}$$

Median lies in the class 30–40

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i = 30 + \frac{75-70}{5} \times 10 = 40$$

$$\text{S.D.}: \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{803}{150} - \left(\frac{59}{150}\right)^2} \times 10$$

$$= \sqrt{5.353 - .155} \times 10 = 2.28 \times 10 = 22.8$$

$$\text{Coeff. of Sk} = \frac{3(38.93-40)}{22.8} = \frac{3(-1.07)}{22.8} = \frac{-3.21}{22.8} = -0.141$$

Illustration 22. Calculate the value of γ_1 and γ_2 from the following data and interpret them:

Profits (Rs. lakhs) :	10–20	20–30	30–40	40–50	50–60
No. of Cos. :	18	20	30	22	10

Solution.

CALCULATION OF β_1 AND β_2

Profits (Rs. lakhs)	*m.p.* X	f	$(X-35)/10$ d	fd	fd^2	fd^3	fd^4
10–20	15	18	−2	−36	72	−144	288
20–30	25	20	−1	−20	20	−20	20
30–40	**35**	30	0	0	0	0	0
40–50	45	22	+1	+22	+22	+22	22
50–60	55	10	+2	+20	+40	+80	160
				$\Sigma fd = -14$	$\Sigma fd^2 = 154$	$\Sigma fd^3 = -62$	$\Sigma fd^4 = 490$

$$\mu'_1 = \frac{\Sigma fd}{N} \times i = \frac{-14}{100} \times 10 = -1.4\,;\ \mu'_2 = \frac{\Sigma fd^2}{N} \times i^2 = \frac{154}{100} \times 100 = 154\,;$$

$$\mu'_3 = \frac{\Sigma fd^2}{N} \times i^3 = \frac{-62}{100} \times 1000 = -620\,;\ \mu'_4 = \frac{\Sigma fd^4}{N} \times i^4 = \frac{490}{100} \times 10000 = 49000$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 152.04 \text{ or } \sigma = \sqrt{\mu_2} = 12.33$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'^3_1 = 21.312$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2(\mu'_1)^2 - (3\mu'_1)^4 = 47327.51$$

$$\gamma_1 = \frac{\mu_3}{\sigma_3} = \frac{21.312}{1874.7140} = 0.0114.$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{47327.51}{(152.04)^2} - 3 = 2.047 - 3 = -0.953.$$

Therefore, $\gamma_1 = 0.0014$ suggests that it is almost near to a symmetrical distribution and γ_2 is less than zero, hence it is a platykurtic curve.

Illustration 23. Calculate Pearson's measure of skewness on the basis of mean, mode and standard deviation, from the following data :

Class-Interval :	14–15	15–16	16–17	17–18	18–19	19–20	20–21	21–22
Frequency :	35	40	48	100	125	87	43	22

(MBA, IGNOU, June 2001)

Solution : CALCULATION OF KARL PEARSON'S COEFFICIENT OF SKEWNESS

Class-Interval	*m.p.* X	f	$(X - 17.5)/1$ d	fd	fd^2
14–15	14.5	35	–3	– 105	315
15–16	15.5	40	–2	– 80	160
16–17	16.5	48	–1	– 48	48
17–18	**17.5**	100	0	0	0
18–19	18.5	125	+1	+ 125	125
19–20	19.5	87	+2	+ 174	348
20–21	20.5	43	+3	+ 129	387
21–22	21.5	22	+4	+ 88	352
		$N = 500$		$\Sigma fd = 283$	$\Sigma fd^2 = 1735$

$$\text{Coeff. of Sk} = \frac{\overline{X} - \text{Mode}}{\sigma}$$

Calculation of Mean : $\overline{X} = A + \frac{\Sigma fd}{N} \times i = 17.5 + \frac{283}{500} \times 1 = 17.5 + 0.57 = 18.07$

Calculation of Standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{1735}{500} - \left(\frac{283}{500}\right)^2} \times i = \sqrt{3.47 - 0.32} = 1.775$$

Calculation of Mode : By inspection mode lies in the class 18–19.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 18, \Delta_1 = f_1 - f_0 = 125 - 100 = 25$$

$$\Delta_2 = f_1 - f_2 = 125 - 87 = 38, i = 1$$

$$\text{Mode} = 18 + \frac{25}{25 + 38} = 18 + .397 = 18.397$$

Substituting the values :

$$\text{Coeff. of Sk} = \frac{18.07 - 18.397}{1.775} = \frac{0.327}{1.775} = 0.184.$$

Illustration 24. The row data displayed below are the observations on the number of passengers who have chosen to fly on Air India in 32 cities, in a particular month.

25	37	23	26	30	40	25	26
39	32	21	26	19	27	32	23
18	26	34	18	31	35	21	33
33	9	16	32	35	42	15	24

(*a*) Construct a frequency distribution using the above data.

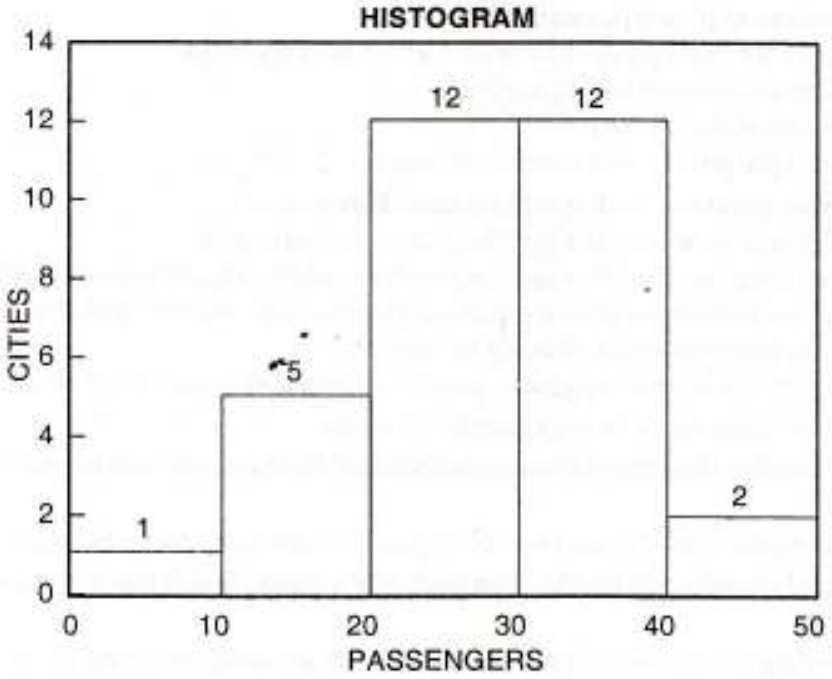
(*b*) Develop and interpret from the above data.

(*c*) Calculate and interpret mean, median, variance and coefficient of variation for the above data.

(*d*) Are the data skewed? Give the coefficient of skewness. *(MBA, Delhi Univ., 2009)*

Solution : PREPARATION OF FREQUENCY DISTRIBUTION

Passengers	*Tally Bars*	*m.p.* m	*Cities* f	$(m-25)/10$ d	fd	fd^2	cf
0–10	|	5	1	–2	–2	4	1
10–20	𝍸	15	5	–1	–5	5	6
20–30	𝍸 𝍸 ||	25	12	0	0	0	18
30–40	𝍸 𝍸 ||	35	12	+1	+12	12	30
40–50	||	45	2	+2	+4	8	32
			$N = 32$		$\Sigma fd = 9$	$\Sigma fd^2 = 29$	



$$\text{Mean}: \bar{X} = A + \frac{\Sigma fd}{N} \times i = 25 + \frac{9}{32} \times 10 = 25 + 2.813 = 27.813$$

$$\text{Median}: \text{Med.} = \text{Size of } \frac{N}{2}\text{th item} = \frac{32}{2} = 16\text{th item}$$

Median lies in the class 20–30

$$\text{Med.} = L + \frac{N/2 - p.c.f.}{f} \times i$$

$$= 20 + \frac{16 - 6}{12} \times 10 = 20 + 8.33 = 28.33$$

$$\text{Standard Deviation}: \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{29}{32} - \left(\frac{9}{32}\right)^2} \times 10 = \sqrt{0.906 - 0.079} \times 10$$

$$= 0.909 \times 10 = 0.09$$

Variance : $\sigma^2 = (9.09)^2 = 82.623$

$$\text{Coeff. of Variation} = \frac{\sigma}{\bar{X}} \times 100 = \frac{9.09}{27.813} \times 100 = 32.68$$

Since it is a bi modal series, skewness will be calculated by formula :

$$\text{Coeff. of Sk} = \frac{3(\bar{X} - \text{Med.})}{\sigma}$$

$\bar{X} = 27.813$, Median = 28.33, $\sigma = 9.09$

$$\text{Coeff. of Sk} = \frac{3(27.813 - 28.33)}{9.09} = \frac{-1.551}{9.09} = -0.171$$

The distribution is skewed to the left. However, there is very low degree of skewness.

PROBLEMS

1-A: Answer the following questions, each question carries **one** mark:

(*i*) What is skewness ?
(*ii*) Point out the role of studying skewness.
(*iii*) Name the various methods of finding skewness.
(*iv*) What are kurtosis ?
(*v*) What are moments ?
(*vi*) How are the values of β_1 and β_2 calculated ?
(*vii*) Give the formula for finding Karl Pearson's coefficient of skewness.
(*viii*) What is Bowley's method of finding skewness ?
(*ix*) What is symmetrical distribution ?
(*x*) Distinguish between positive and negative skewness .

1-B: Answer the following questions, each question carries **four** marks:

(*i*) Distinguish between positively and negatively skewed distribution.
(*ii*) In what type of situations Karl Pearson's or Bowley's method should be preferred ?
(*iii*) Would the various methods of studying skewness lead to same answer ? If not, give reasons.
(*iv*) What are the various methods of studying kurtosis ?
(*v*) Explain the terms leptokurtic, platykurtic and mesokurtic with a suitable diagram.

2. (*a*) Explain briefly the different methods of measuring skewness.
(*b*) What do you understand by the terms skewness and kurtosis? Point out their role in analysing a frequency distributi
(MBA, Delhi Univ., 20

3. Take any suitable imaginary data and explain how would you measure skewness and kurtosis.

4. Distinguish between Karl Pearson's and Bowley's measure of skewness. Which one of these would you prefer and wh
(MBA, Delhi Univ., 2

5. Measures of central, tendency, variation, skewness, and kurtosis are complementary to one another in understandi frequency distribution? Elucidate. *(MBA, Sukhadia Univ.; Delhi Univ., 2*

6. Define 'Moments'. How can you find out skewness and kurtosis of a distribution from moments about the mean?

7. Explain clearly how the moments help in describing the characteristics of a frequency distribution.

8. Explain clearly how the measures of skewness and kurtosis can be used in describing a frequency distribution.

9. What is meant by 'moments' of a distribution ? Show how moments are used to describe the characteristics of a distrib *i.e.*, central tendency, dispersion, skewness and kurtosis.

10. What are the raw and the central moments of a distribution? Show that the central moments are invariant under cha origin but not under change of scale.

11. Define raw and central moments of a frequency distribution. Express the second, third and fourth central moments ir of raw moments.

12. (*a*) Explain the terms 'Skewness' and 'Kurtosis' used in connection with the frequency distribution of a continuous v Give the different measures of skewness (any two of the measures to be given) and kurtosis.
(*b*) Define and discuss the 'quartiles' of a distribution. How are they used for measuring variation and skewness

13. Define moments. Establish the relationship between the moments about mean in terms of moments about any arbitrary point and *vice-versa*.

14. (*a*) Define moments. How are they helpful in study of the different aspects of the formation of a frequency distribution?
(*b*) "A frequency distribution can be described almost completely by the first four moments and the two measures based on the moments." Examine.

15. (*a*) Explain the third and fourth central moment in terms of the first four moments about the origin.
(*b*) Distinguish between variation and skewness and point out the various methods of measuring skewness.
(*c*) Explain the term 'skewness'. What purpose does a measure of skewness serve? Comment on some of the well-known measures of skewness.

16. (*a*) Distinguish between skewness and kurtosis.
(*b*) Briefly mention the tests which can be applied to determine the presence of skewness.

17. (*a*) How do measures of central tendency, dispersion, skewness and kurtosis help in analysing a frequency distribution? Explain with the help of an example. *(MBA, Sukhadia Univ., 2008)*
(*b*) Find out coefficient of skewness from the following table giving wages of 240 persons :

Wages (Rs.)	*No. of persons*	*Wages (Rs.)*	*No. of persons*
2000–2200	12	2800–3000	50
2200–2400	18	3000–3200	45
2400–2600	35	3200–3400	30
2600–2800	42	3400–3600	8

[Sk = – 0.267]

18. Calculate Karl Pearson's coefficient of skewness from the following data :

Profits (Rs. Lakhs)	*No. of Cos.*	*Profits (Rs. Lakhs)*	*No. of Cos.*
400–450	8	600–650	62
450–500	10	650–700	32
500–550	30	700–750	15
550–600	45	750–800	8

19. The following data represent the percentage of ash content in a particular variety of coal as determined by test on 280 wagon loads :

Percentage of ash content	*Frequency*	*Percentage of ash content*	*Frequency*
Less than 6.0	0	10.0 –10.9	84
6.0– 6.9	1	11.0 –11.9	45
7.0– 7.9	7	12.0–12.9	28
8.0– 8.9	28	13.0–13.9	7
9.0– 9.9	78	14.0–14.9	2

Calculate the quartile coefficient of skewness. Also compare the proportion of the total frequency outside the limits $\bar{X} \pm 2\sigma$ for the distribution.

[Sk=0.05; 2.3]

20. From the following data of daily travelling allowance (in Rs.) of salesmen, calculate coefficient of skewness and comment on its value :

Travelling allowance (per day)	*No. of salesmen*	*Travelling allowance (per day)*	*No. of salesmen*
110–115	4	135–140	90
115–120	10	140–145	52
120–125	26	145–150	33
125–130	49	150–155	17
130–135	72	155–160	7

21. From the following data pertaining to profits (Rs. lakhs) for 50 companies, calculate moments β_1 and β_2 :

Profits (Rs. Lakhs)	*No. of Compaincs*
70–90	8
90–110	11
110–130	18
130–150	9
150–170	4

[$\mu_2 = 528$, $\mu_3 = 960$, $\mu_4 = 642816$, $\beta_1 = 0.006$, $\beta_2 = 2.31$]

22. A record was kept over a period of 6 months by a sales manager to determine the average number of calls made per day by his six salesmen. The results are shown below :

Salesmen	:	*A*	*B*	*C*	*D*	*E*	*F*
Average number of calls per day	:	8	10	12	15	7	5

(*i*) Compute a measure of skewness. Is the distribution symmetrical ?

(*ii*) Compute a measure of kurtosis. What does this measure mean ?

[β_1=0.11; β_2=1.97]

23. Locate the mode and calculate mean and standard deviation of the following distribution and using your results comment on the skewness of the distribution :

Scores	*Frequency*	*Scores*	*Frequency*
10–15	2	35–40	6
15–20	8	40–45	4
20–25	6	45–50	3
25–30	12	50–55	1
30–35	7	55–60	1

[$\bar{X}$ = 30.1; Mo.= 27.73,σ = 10.45, Sk = 0.227]. (*MBA, Delhi Univ., 2002, 2005*)

24. You are given the following information before and after the settlement of an industrial dispute :

	Before settlement of dispute	*After settlement of dispute*
No. of workers	1100	950
Average wage (Rs.)	2350	2400
Standard deviation (Rs.)	425	400
Median wage (Rs.)	2375	2325

Comment on the gains and losses from the point of view of workers and that of management.

25. The arithmetic mean of a distribution is 5. The second and the third moments about the mean are 20 and 140 respectively Find the third moment of the distribution about 10.

[–285]

26. For the frequency distribution given below, calculate the coefficient of skewness based on the quartiles :

Class limits	*Frequency*	*Class limits*	*Frequency*
10–19	5	50–59	25
20–29	9	60–69	5
30–39	14	70–79	8
40–49	20	80–89	4

27. (*a*) For a distribution, Bowley's coefficient of skewness is –0.48, Q_3 = 10.2 and Median = 14.4. What is the quarti coefficient of distribution?

(*b*) Karl Pearson's coefficient of skewness of a distribution is +0.4. Its standard deviation is 10 and mean 40.5. Find tl mode and median of the distribution.

(*c*) Find coefficient of skewness from the information given below :

Q_1 = 60, Q_3 = 75, Med. = 68.

(*d*) The following information was obtained from the records of a factory relating to wages; $\bar{X}$ = 275, Med. = 260, σ 45.8

Give as much information as you can about the distribution of wages.

[(*a*) 0.22 (*b*) 39.17 (*c*) –0.07 (*d*) Sk = 0.98]

28. The first three moments of a distribution about the value 7 calculated from a set of 9 observations are 0.2, 19.4 and –41 Find the measures of central tendency and dispersion and also the third moment about origin.

[7.2, 4.4, –52.624]

29. The first four moments of a distribution about A = 4 are 1, 4, 10 and 45. Obtain the various characteristics of the distribut on the basis of the information given. Comment upon the nature of the distribution.

[β_1 = 0, β_2 = 2.897]

30. (*a*) State the use of quartiles for measuring dispersion and skewness.

(*b*) Calculate Bowley's coefficient of skewness from the following data :

Mid-value	:	75	100	125	150	175	200	225	250
Frequency	:	35	40	48	100	125	80	50	22

[–0.032]

31. A prospective buyer tested the bursting pressure of the sample of polythene bags received from a manufacturer. The test gives the following results :

Bursting pressure (in lbs.)	5–10	10–15	15–20	20–25	25–30	30–35
No. of bags	2	20	30	50	6	2

The buyer calculated the mean and mode of the sample as 20.2 lbs. and 21.5 lbs. respectively.
Calculate (*i*) coefficient of variation, (*ii*) Karl Pearson's coefficient of skewness for bursting pressure.

32. From the following data, calculate coefficient of variation and coefficient of skewness :

Age (in years) :	25–30	30–35	35–40	40–45	45–50	50–55	55–60
No. of employees :	9	18	30	40	10	7	6

33. The frequency distribution of weekly wages (in Rs.) in a certain factory is as follows :

Weekly wages	*No. of workers*	*Weekly wages*	*No. of workers*
423–427	2	448–452	16
428–432	6	453–457	12
433–437	9	458–462	6
438–442	14	463–467	2
443–447	32	468–472	1

Find Karl Pearson's coefficient of skewness and interpret its value.
[$Sk_P = 0.0572$]

34. A survey was conducted by a manufacturing company to enquire the maximum price at which persons would be willing to buy their product. The following table gives the stated prices (in rupees) by persons :

Price (in Rs.) :	80–90	90–100	100–110	110–120	120–130
No. of persons :	11	29	18	27	15

Calculate Bowley's coefficient of skewness and interpret its value. *(MBA, Delhi Univ., 2002)*

35. The standard deviation of a symmetrical distribution is 3. What must be the value of fourth moment about the mean in order that the distribution be mesokurtic?

36. Calculate coefficient of variation and Karl Pearson's coefficient of skewness from the data given below :

Sales (Rs. crores) Less than :	40	50	60	70	80
No. of Companies :	8	20	50	72	80

[Coeff. of Variation = 19.55, Coeff. of Sk = –0.06]

37. Assume that a firm has selected a random sample of 100 from its production line and has obtained 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	Total	100

Compute Karl Pearson's Coefficient of Skewness.
[Coeff. of Sk = –0.572]

(MBA, Mangalore Univ., 2005)

38. (*a*) A moderately skewed distribution has mean and median as 25 and 26 respectively. Then its mode approximately equals.....

(*b*) Whether the following statement is *true* or *false* : If a distribution has negative skewness then its mean is greater than mode.

39. Calculate the first four moments about mean and find the values of β_1 and β_2 and comment on the result :

Profits (Rs. lakhs) :	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No. of Companies :	8	12	20	30	15	10	5

(MBA, Kumaun Univ., 2004)

40. From the following data pertaining to the income of 5,800 persons, find Bowley's coefficient of skewness and interpret its value :

Income (Rs.)	*No. of persons*	*Income (Rs.)*	*No. of persons*
Below 10,000	170	40,000–50,000	1,350
10,000–20,000	630	50,000–60,000	1,000
20,000–30,000	1,000	60,000 and above	400
30,000–40,000	1,250		

[Coeff. of Sk = –0.067]

(MBA, Kurukshetra Univ., 2001)

41. Compute the first 3 moments about the arithmetic mean from the following data :

Variable value :	5	10	15	20	25	30	35
Frequency :	8	15	20	32	23	17	5

(MBA, Lucknow Univ., 2001)

42. The following distribution gives the pattern of overtime work done in a month by 100 employees of a company :

Overtime hours :	10–15	15–20	20–25	25–30	30–35	35–40
No. of employees :	11	20	35	20	8	6

Compute mean, mode, standard deviation and coefficient of skewness.

[23.1, 22.5, 6.4915, 0.0924]

43. The following table gives the distribution of monthly wages of 500 workers in a factory:

Monthly wages (Rs. hundred)	*No. of workers*	*Monthly wages (Rs. hundred)*	*No. of workers*
15–20	10	30–35	220
20–25	25	35–40	70
25–30	145	40–45	30

Compute Karl Pearson's and Bowley's coefficient of skewness. Interpret the values.

[$SK_P = -0.022$, $SK_B = -0.102$]

(MBA, Delhi Univ., 2006)

44. Calculate Karl Pearson's coefficient of skewness from the data given below :

Marks	*No. of candidates*	*Marks*	*No. of candidates*
70–80	11	30–40	21
60–70	22	20–30	11
50–60	30	10–20	6
40–50	35	0–10	5

[–0.026]

(MBA, Kumaun Univ., 2001)

45. Calculate β_1 and β_2 from the following distribution and interpret the results :

Age	*Frequency*	*Age*	*Frequency*
25–30	2	45–50	25
30–35	8	50–55	16
35–40	18	55–60	7
40–45	27	60–65	2

[$\beta_1 = 0.034$, $\beta_2 = 2.59$]

*****

7

Correlation Analysis

INTRODUCTION

So far we have studied problems relating to one variable only. In business we come across a large number of problems involving the use of two or more than two variables. If two quantities vary in such a way that movements in one are accompanied by movements in the other, these quantities are said to be correlated. For example, there exists some relationship between family income and expenditure on luxury items, price of a commodity and amount demanded, increase in rainfall up to a point and production of rice, an increase in the number of television licences and number of cinema admissions, etc. The statistical tool with the help of which these relationships between two or more than two variables is studied is called **correlation***. The measure of correlation called the coefficient of correlation (denoted by the symbol r) summarizes in one figure the direction and degree of correlation. Thus correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables. A very simple definition of correlation is that given by A.M. Tuttle. He defines correlation as : "An analysis of the covariation of two or more variables is usually called *correlation*."

The problem of analysing the relation between different series should be broken down into three steps :

(1) Determining whether a relation exists and, if it does, measuring it;

(2) Testing whether it is significant; and

(3) Establishing the cause-and-effect relations, if any.

In this chapter only the first aspect will be discussed. For second aspect a reference may be made to chapter on Tests on Hypothesis. The third aspect in the analysis, that of establishing the cause-effect relation, is beyond the scope of this text. An extremely high and significant correlation between the increase in smoking and increase in lung cancer would not prove that smoking causes lung cancer.

It should be noted that the detection and analysis of correlation (*i.e.*, convariation) between two statistical variables requires relationship of some sort which associates the observation in pairs, one of each pair being a value of each of the two variables. In general, the pairing relationship may be of almost any nature, such as observations at the same time or place or over a period of time or different places.

Significance of the Study of Correlation

The study of correlation is of immense use in practical life because of the following reasons :

1. Most of the variables show some kind of relationship between price and supply, income and expenditure, etc. With the help of correlation analysis we can measure in one figure the degree of relationship existing between the variables.

---

*"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."—Croxton and Cowden : *Applied General Statistics.*

2. Once we know that two variables are closely related, we can estimate the value of one variable given the value of another. This is done with the help of regression analysis which is discussed in the next chapter.

3. Correlation analysis contributes to the economic behaviour, aids in locating the critically important variables on which others depend, *may reveal to the economist the connection by which disturbances spread and suggest to him the paths through which stabilising forces become effective.*

In business, correlation analysis enables the executive to estimate costs, sales, price and other variables *on the basis of some other series with which these costs, sales, or prices may be functionally related.* Some of the guesswork can be removed from decisions when the relationship between a variable to be estimated and the one or more other variables on which it depends are close and reasonably invariant.

4. Progressive development in the methods of science and philosophy has been characterised by increase in the knowledge of relationship or correlations. Nature has been found to be multiplicity of *inter-related forces.*

However, it should be noted that coefficient of correlation is one of the most widely used and also one of the most widely *abused statistical measures. It is abused in the sense* that one sometimes overlooks the fact that correlation measures nothing but the strength of *linear* relationships and that it does not necessarily imply a relationship.

Correlation and Causation

Correlation analysis helps us in determining the degree of relationship between two or more variables—it does not tell us anything about cause-effect relationship. Even a high degree of correlation does not necessarily mean that a relationship of cause and effect exists between the variables or, simply stated, correlation does not necessarily imply causation or functional relationship though the existence of causation always implies correlation. By itself it establishes only *covariation.* The explanation of significant degree of correlation may be any one, or a combination of the following factors :

1. *The correlation may be due to pure chance, especially in a small sample.* We may get a high degree of correlation between two variables in the sample but in the universe, there may not be any relationship between the variables at all. This is especially so in case of small samples. Such a correlation may arise either because of pure random sampling variation or because of the bias of the investigator in selecting the sample. The following example shall illustrate the point :

Advertisement expenditure (Rs. lakhs)	*Sales (Rs. crores)*
25	120
35	140
45	160
55	180
65	200

The above data show a perfect positive relationship between advertisement expenditure and sales, *i.e.*, as the advertisement expenditure is increasing, the sales are also increasing and the ratio of change between the two variables is the same. However, such a situation is rare in practice.

2. *Both the correlated variables may be influenced by one or more other variables.* It is just possible that a high degree of correlation between the variables may be due to the same causes affecting each variable or different causes affecting each with the same effect. For example, a high degree of correlation between the yield per acre of rice and tea may be due to the fact that both are related to the amount of rainfall. But none of the two variables is the cause of the other.

3. *Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect.* There may be a high degree of correlation between the variables but it may be difficult to pinpoint as to which is the cause and which is the effect. This is especially likely to be so in case of economic variables. For example, such variables as demand and supply, price and production, etc., mutually interact. To take a specific case, it is a well-known principle of economics that as the price of a commodity increases, its demand goes down and so price is the cause and demand the effect. But it is also possible that increased demand of a commodity due to growth of population or other reasons may force its price up. Now the cause is the increased demand, the effect the price. Thus at times it may become difficult to explain from the two correlated variables which is the cause and which is the effect because both may be reacting on each other.

The above points clearly bring out the fact that correlation does not manifest causation or functional relationship. By itself, it establishes only covariation. Correlation observed between variables that could not conceivably be causally related are called *spurious or nonsense correlation*. More appropriately, we should remember that it is the *interpretation* of the degree of correlation that is spurious, not the degree of correlation itself. The high degree of correlation indicates only the mathematical result. We should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. A last word of warning: Errors in correlation analysis include not only reading causation into spurious correlation but also interpreting spuriously a perfectly valid association.

Types of Correlation

Correlation is described or classified in several different ways. Three of the most important are :

(*i*) Positive and negative ;
(*ii*) Simple, partial and multiple ; and
(*iii*) Linear and non-linear.

(*i*) **Positive and Negative Correlation.** Whether correlation is positive (direct) or negative (inverse) would depend upon the direction of change of the variable. If both the variables are varying in the same direction, *i.e.*, if one variable is increasing the other *on an average* is also increasing or, if one variable is decreasing the other *on an average* is also decreasing, correlation is said to be positive. If, on the other hand, the variables are varying in opposite directions, *i.e.*, as one variable is increasing the other is decreasing or *vice versa*, correlation is said to be negative. The following examples would illustrate positive and negative correlation :

POSITIVE CORRELATION

X	Y
10	15
12	20
11	22
18	25
20	37

POSITIVE CORRELATION

X	Y
80	50
70	45
60	30
40	20
30	10

NEGATIVE CORRELATION

X	Y
20	40
30	30
40	22
60	15
80	16

NEGATIVE CORRELATION

X	Y
100	10
90	20
60	30
40	40
30	50

(*ii*) **Simple, Partial and Multiple Correlation.** The distinction between simple, partial and multiple correlation is based upon the number of variables studied. When only two variables are studied it is a problem of simple correlation. When three or more variables are studied it is a problem of either multiple or partial correlation. In multiple correlation three or more variables are studied simultaneously. For example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilisers used, it is a problem of multiple correlation. Similarly, the relationship of plastic hardness, temperature and pressure is multivariate. In partial correlation we recognise more than two variables. But consider only two variables to be influencing each other, the effect of other influencing variable being kept constant. For example, in the rice problem taken above if we limit our correlation analysis of yield and rainfall to periods when a certain average daily temperature existed, it becomes a problem of partial correlation. In this chapter, we shall study problems relating to simple correlation only.

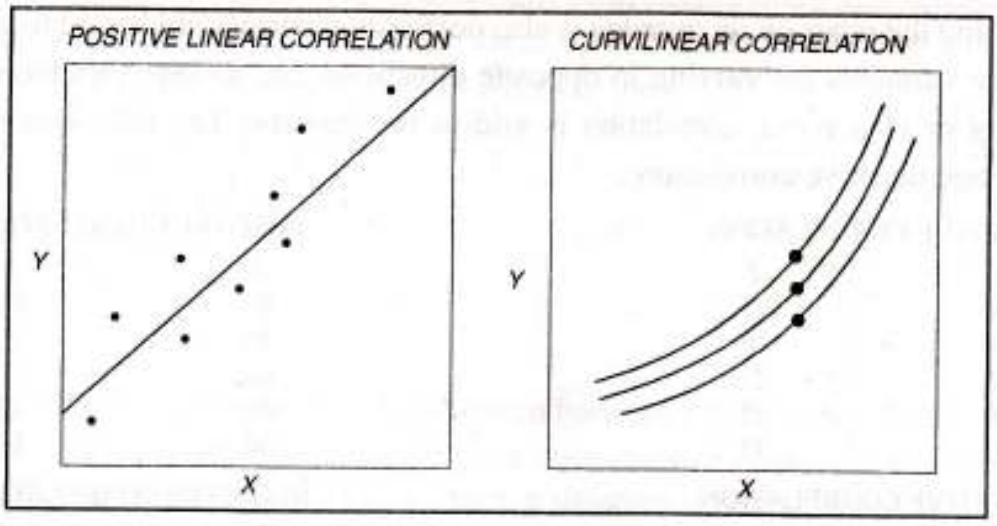
(*iii*) **Linear and Non-linear (Curvilinear) Correlation.** The distinction between linear and non-linear correlation is based upon the constancy of the ratio of change between the variables. If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. For example, observe the following two variables X and Y :

X:	10	20	30	40	50
Y:	70	140	210	280	350

It is clear that the ratio of change between the two variables is the same. If such variables are plotted on a graph paper, all the plotted points would fall on a straight line.

Correlation would be called non-linear or curvilinear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. For example, if we double the amount of rainfall, the production of rice or wheat, etc., would not necessarily be doubled. It may be pointed out that in most practical cases we find a non-linear relationship between the variables. However, since techniques of analysis for measuring non-linear correlation are far more complicated than those for linear correlation, we generally make an assumption that the relationship between the variables is of the linear type.

The following two diagrams will illustrate the difference between linear and curvilinear correlation :



METHODS OF STUDYING CORRELATION

The following are the important methods of ascertaining whether two variables are correlated or not :

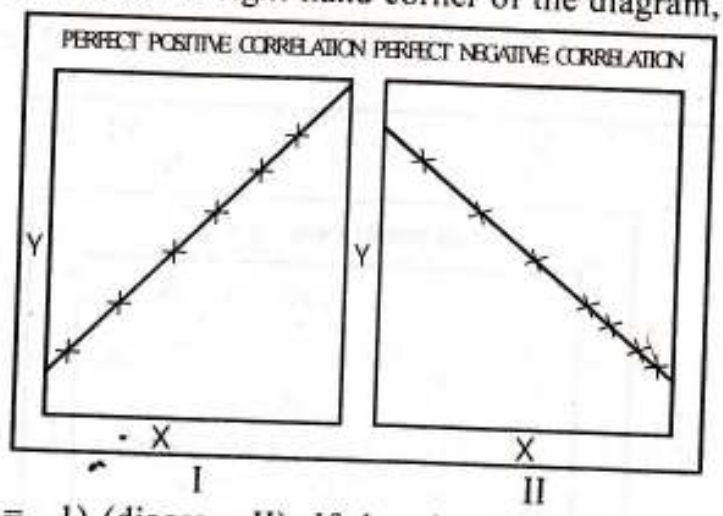
I. Scatter Diagram Method ;

II. Karl Pearson's Coefficient of Correlation ;
III. Spearman's Rank Correlation Coefficient ; and
IV. Method of Least Squares.*

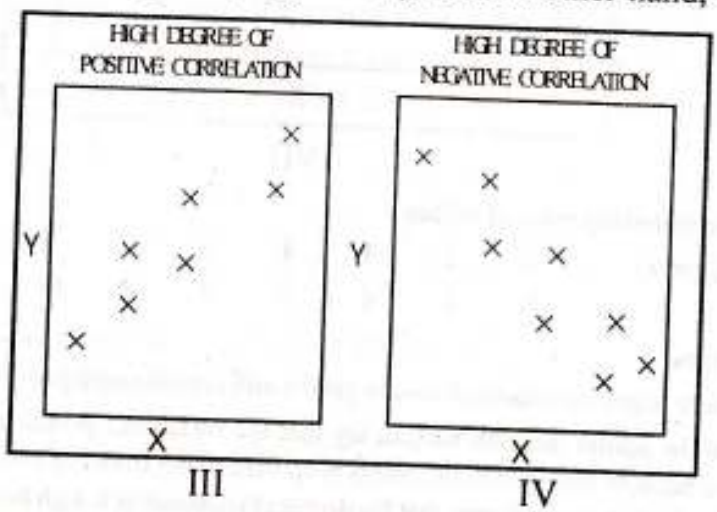
Of these, the first one is based on the knowledge of graphs whereas the others are the mathematical methods. Each of these methods shall be discussed in detail in the following pages.

I. SCATTER DIAGRAM METHOD

The simplest device for studying correlation in two variables is a special type of dot chart called dotogram or scatter diagram. When this method is used, the given data are plotted on a graph paper in the form of dots, *i.e.*, for each pair of X and Y values we put dots and thus obtain as many points as the number of observations. By looking to the scatter of the various points, we can form an idea as to whether the variables are related or not. The more the plotted points "scatter" over a chart, the lesser is the degree of relationship in between the two variables. The more nearly the points come to the line, the higher the degree of relationship. If all the points lie on a straight line falling from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfectly positive (*i.e.*, $r = +1$) (diagram I). On the other hand, if all the points are lying on a straight line rising from the upper left-hand corner to the lower right-hand corner of the diagram, correlation is said to be

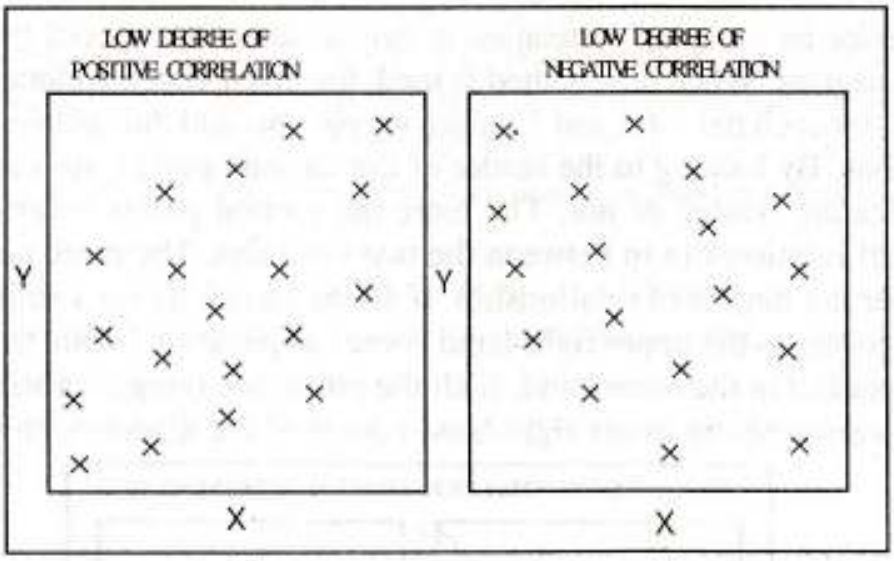


perfectly negative (*i.e.*, $r = -1$) (diagram II). If the plotted points fall in a narrow band, there would be a high degree of correlation between the variables—correlation shall be positive if the points show a rising tendency from the lower left-hand corner to the upper right-hand corner (diagram III) and negative if the points show a declining tendency from upper left-hand corner to the lower right-hand corner of the diagram (diagram IV). On the other hand, if the points are widely

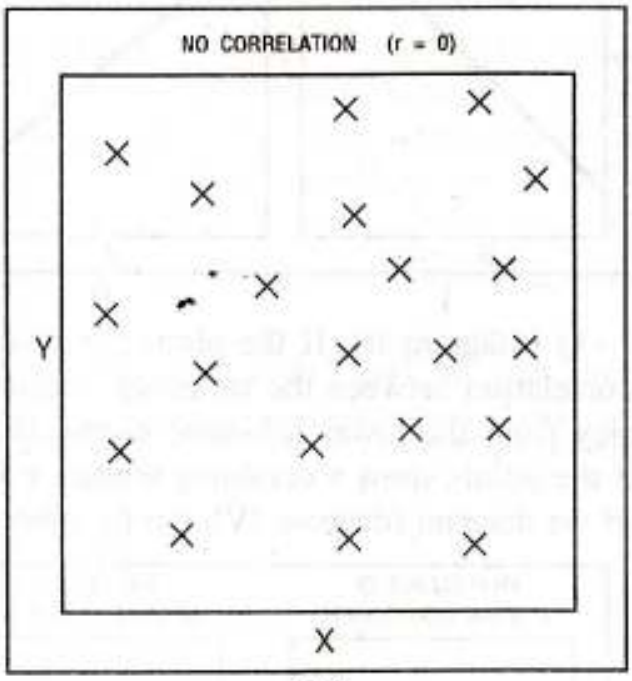


*This method is discussed in detail in Chapter on 'Regression Analysis'.

scattered over the diagrams it indicates very low degree of relationship between the variables—correlation shall be positive if the points are rising from the lower left-hand corner to the upper right-hand corner (diagram V) and negative if the points are running from the upper left-hand side to the lower right-hand side to the diagram (diagram VI). If the plotted points lie on a straight line parallel to the X-axis, or in a haphazard manner, it shows the absence of any relationship between the variables (*i.e.*, $r = 0$) as shown by diagram VII.



V VI



VII

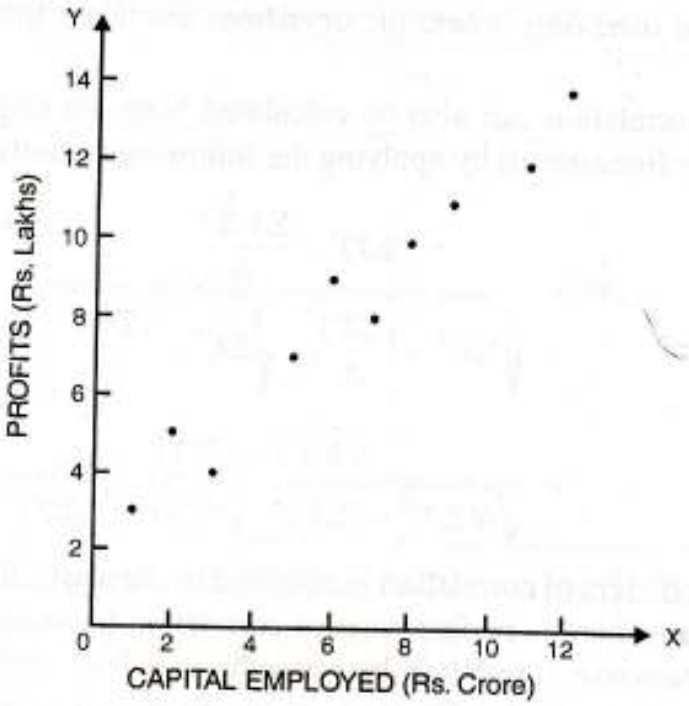
Illustration 1. Given the following pairs of values :

Capital employed (Rs. Crore) :	1	2	3	4	5	7	8	9	11	12
Profits (Rs. Lakhs) :	3	5	4	7	9	8	10	11	12	14

(*a*) Make a scatter diagram.

(*b*) Do you think that there is any correlation between profits and capital employed ? Is it positive ? Is it high or low ?

Solution. By looking at the scatter diagram we can say that the variables : profits and capital employed are correlated. Further, correlation is positive because the trend to the points is upward rising from the lower left-hand corner to the upper right-hand corner of the diagram. The diagram also indicates that the degree of relationship is high because the plotted points are in a narrow band which shows that it is a case of high degree of positive correlation.



Merits and Limitations of the Method

Merits : 1. It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.

2. It is not influenced by the size of extreme values whereas most of the mathematical methods of finding correlation are influenced by extreme values.

3. Making a scatter diagram usually is the first step in investigating the relationship between the variables.

Limitations. By applying this method we can get an idea about the direction of correlation and also whether it is high or low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical method.

II. KARL PEARSON'S COEFFICIENT OF CORRELATION

Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice. The coefficient of correlation is denoted by the symbol r. It is one of the very few symbols that is used universally for describing the degree and direction of relationship between two variables. If the two variables under study are X and Y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship.

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\sqrt{\Sigma(Y - \bar{Y})^2}} \quad ...(i)$$

where $\bar{X}$ and $\bar{Y}$ are the respective means of X and Y variable.

The above formula can be written as :

$$r^* = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \quad ...(ii)$$

where $x = (X - \bar{X})$ and $y = (Y - \bar{Y})$.

This formula is to be used only where the deviations are taken from *actual* means and *not* from assumed means.

The coefficient of correlation can also be calculated from the original set of observations (*i.e.*, without taking deviations from mean) by applying the following formula :

$$r^{**} = \frac{\Sigma XY - \frac{\Sigma X\, \Sigma Y}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\ \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

$$= \frac{N\Sigma XY - \Sigma X\, \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \qquad ...(iii)$$

The value of the coefficient of correlation as obtained by the above formula shall always lie between ± 1. When $r = +1$, it means there is perfect positive correlation between the variables. When $r = -1$, it means there is perfect negative correlation between the variables. When $r = 0$, it means there is no relationship between the two variables. However, in practice, such value of r as $+1$, -1, and 0 are rare. We normally get values which lie between $+1$ and -1 such as 0.8, -0.4, etc. The coefficient of correlation describes not only the magnitude of correlation but also its direction. Thus, $+0.8$ would mean that correlation is positive because the sign of r is +ve and the magnitude of correlation is 0.8.

The following illustration will clarify the procedure of computing the coefficient of correlation :

Illustration 2. Find correlation coefficient between the sales and expenses from the data given below :

Firm	:	1	2	3	4	5	6	7	8	9	10
Sales (Rs. Lakhs)	:	50	50	55	60	65	65	65	60	60	50
Expenses (Rs. Lakhs)	:	11	13	14	16	16	15	15	14	13	13

---

*The coefficient of correlation can also be expressed in terms of covariance and variance as given below :
From (*ii*), we have

$$r = \frac{\Sigma xy / N}{\sqrt{\Sigma x^2 / N}\ \sqrt{\Sigma y^2 / N}} = \frac{\text{Cov}\,[x, y]}{\sqrt{\text{Var}\ x, \text{Var}\ y}} = \frac{\text{Cov}\,[x, y]}{\sigma_x\, \sigma_y}$$

**This formula is derived from formula (*i*) as follows :

$$r = \frac{\Sigma(X - \bar{X})\, \Sigma(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\ \sqrt{\Sigma(Y - \bar{Y})^2}}$$

Opening the brackets, we get :

$$r = \frac{\Sigma XY - N\bar{X}\,\bar{Y}}{\sqrt{\Sigma X^2 - N\bar{X}^2}\ \sqrt{\Sigma Y^2 - N\bar{Y}^2}}$$

$$= \frac{\Sigma XY - \frac{\Sigma X . \Sigma Y}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\ \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

$$= \frac{N\Sigma XY - \Sigma X\, \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Solution. CALCULATION OF CORRELATION COEFFICIENT

Firm	*Sales* X	$(X-\bar{X})$ x	x^2	*Expenses* Y	$(Y-\bar{Y})$ y	y^2	xy
1	50	–8	64	11	–3	9	+24
2	50	–8	64	13	–1	1	+8
3	55	–3	9	14	0	0	0
4	60	+2	4	16	+2	4	+4
5	65	+7	49	16	+2	4	+14
6	65	+7	49	15	+1	1	+7
7	65	+7	49	15	+1	1	+7
8	60	+2	4	14	0	0	0
9	60	+2	4	13	–1	1	–2
10	50	–8	64	13	–1	1	+8
$N=10$	$\Sigma X=580$	$\Sigma x=0$	$\Sigma x^2=360$	$\Sigma Y=140$	$\Sigma y=0$	$\Sigma y^2=22$	$\Sigma xy=70$

$$\bar{X}=\frac{\Sigma X}{N}=\frac{580}{10}=58;\ \bar{Y}=\frac{\Sigma Y}{N}=\frac{140}{10}=14$$

$$r=\frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}=\frac{70}{\sqrt{360\times 22}}=\frac{70}{88.994}=0.787$$

To simplify calculation we can solve the question with the help of logarithms also.
Taking logarithms

$$\log r = \log 70 - \tfrac{1}{2}(\log 360 + \log 22)$$
$$= 1.8451 - \tfrac{1}{2}(2.5563 + 1.3424)$$
$$= 1.8451 - \tfrac{1}{2}(3.8987) = 1.8451 - 1.9493 = \bar{1}.8958$$
$$r = AL\ \bar{1}.8958 = 0.787$$

Hence, there is a high degree of positive correlation between the two variables *i.e.*, as the value of sales goes up, the expenses also go up.

When Deviations are taken from an Assumed Mean

When actual means are in fractions, say the actual means of X and Y series are 20.167 and 29.23, the calculation of coefficient of correlation by the method discussed above would involve too many calculations and would take a lot of time. In such cases we make use of the assumed mean method for finding out coefficient of correlation. When deviations are taken from an assumed mean, the following is applicable :

$$r=\frac{N\Sigma d_x d_y-(\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2-(\Sigma d_x)^2}\sqrt{N\Sigma d_y^2-(\Sigma d_y)^2}} \quad \ldots (iv)$$

where d_x refers to deviations of X series from an assumed mean, *i.e.*, $(X–A)$.

Similarly, d_y refers to deviation of Y series from an assumed mean *i.e.*, $(Y–A)$.

It may be noted that this form of formula is same as (*iii*), only difference being that whereas in form (*iii*) we are dealing with original X and Y, in form (*iv*) we are taking deviations of X and Y series from assumed mean.

The following example shall illustrate the application of this formula :

Illustration 3. The following data relate to the age of 10 employees and the number of days which they reported sick in a month :

Age	:	20	30	32	35	40	46	52	55	58	62
Sick days	:	11	12	10	13	14	16	15	17	18	19

Calculate Karl Pearson's coefficient of correlation and interpret its value.

(MBA, DU, 2003)

Solution. Let age and sick days be represented by variable X and Y respectively.

CALCULATION OF CORRELATION COEFFICIENT

Age X	$(X-43)$ d_x	d_x^2	*Sick days* Y	$(Y-14)$ d_y	d_y^2	$d_x d_y$
20	–23	529	11	– 3	9	+ 69
30	– 13	169	12	– 2	4	+ 26
32	– 11	121	10	– 4	16	+ 44
35	– 8	64	13	– 1	1	+ 8
40	– 3	9	14	0	0	0
46	+ 3	9	16	+ 2	4	+ 6
52	+ 9	81	15	+ 1	1	+ 9
55	+ 12	144	17	+ 3	9	+ 36
58	+ 15	225	18	+ 4	16	+ 60
62	+ 19	361	19	+ 5	25	+ 95
$\Sigma X = 430$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 1712$	$\Sigma Y = 145$	$\Sigma d_y = 5$	$\Sigma d_y^2 = 85$	$\Sigma d_x d_y = 353$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{10\times 353 - (0)(-5)}{\sqrt{10\times 1712 - (0)^2}\sqrt{10\times 85 - (-5)^2}}$$

$$= \frac{3530}{\sqrt{17120}\sqrt{825}} = \frac{3530}{130.85\times 28.72} = 0.939$$

Thus, there is a very high degree of positive correlation between age and sick days taken. Hence, we can conclude that as the age of an employee increases, he is liable to be sick more often than others.

Illustration 4. Find the coefficient of correlation by Karl Pearson's method between X and Y and interpret its value.

X	:	57	42	40	33	42	45	42	44	40	56	44	43
Y	:	10	60	30	41	29	27	27	19	18	19	31	29

(MBA, M.D. Univ., 2007)

Solution. CALCULATION OF KARL PEARSON'S CORRELATION COEFFICIENT

X	$(X-44)$ d_x	d_x^2	Y	$(Y-30)$ d_y	d_y^2	$d_x d_y$
57	+ 13	169	10	– 20	400	– 260
42	– 2	4	60	+ 30	900	– 60
40	– 4	16	30	0	0	0
33	– 11	121	41	+ 11	121	– 121
42	– 2	4	29	– 1	1	+ 2
45	+ 1	1	27	– 3	9	– 3
42	– 2	4	27	– 3	9	+ 6
44	0	0	19	– 11	121	0
40	– 4	16	18	– 12	144	+ 48
56	+ 12	144	19	– 11	121	– 132
44	0	0	31	+ 1	1	0
43	– 1	1	29	– 1	1	+ 1
$\Sigma X = 528$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 480$	$\Sigma Y = 340$	$\Sigma d_y = -20$	$\Sigma d_y^2 = 1828$	$\Sigma d_x d_y = -519$

$$r = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{12(-519) - (0)(-20)}{\sqrt{12(480) - (0)^2}\sqrt{12(1828) - (-20)^2}}$$

$$= \frac{-6228}{\sqrt{5760}\sqrt{21536}} = \frac{-6228}{11137.66} = -0.559.$$

Therefore, it is a case of moderate degree of negative correlation.

Correlation of Bivariate Grouped Data

When we have to find coefficient of correlation from a bivariate grouped data table, the following formula is applicable :

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

This formula is the same as that of *(iv)*. The only difference is that here the deviations are also multiplied by the frequencies.

The following illustration shall explain the application of this formula :

Illustration 5. Find the coefficient of correlation between the age and the sum assured from the following table :

Sum assured (in Rs.)

Age group	10,000	20,000	30,000	40,000	50,000
20–30	4	6	3	7	1
30–40	2	8	15	7	1
40–50	3	9	12	6	2
50–60	8	4	2	—	—

(MBA, Delhi Univ., 1999)

Solution. Let the sum assured be denoted by X and the age group by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

Y	X		10,000	20,000	30,000	40,000	50,000				
		d_x / d_y	−2	−1	0	1	2	f	fd_y	fd_y^2	fd_xd_y
20–30	m.p. 25	−2	4 (16)	6 (12)	3 (0)	7 (−14)	1 (−4)	21	−42	84	+10
30–40	35	−1	2 (4)	8 (8)	15 (0)	7 (−7)	1 (−2)	33	−33	33	+3
40–50	45	0	3 (0)	9 (0)	12 (0)	6 (0)	2 (0)	32	0	0	0
50–60	55	+1	8 (−16)	4 (−4)	2 (0)	—	—	14	+14	14	−20
		f	17	27	32	20	4	$N = 100$	$\Sigma fd_y = -61$	$\Sigma fd_y^2 = 131$	$\Sigma fd_xd_y = -7$
		fd_x	−34	−27	0	20	−8	$\Sigma fd_x = -33$			
		fd_x^2	68	27	0	20	16	$\Sigma fd_x^2 = 131$			
		fd_xd_y	4	16	0	−21	−6	$\Sigma fd_xd_y = -7$			

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{100(-7) - (-33)(-61)}{\sqrt{100(131) - (-33)^2}\sqrt{100(131) - (-61)^2}}$$

$$r = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{12(-519) - (0)(-20)}{\sqrt{12(480) - (0)^2}\sqrt{12(1828) - (-20)^2}}$$

$$= \frac{-6228}{\sqrt{5760}\sqrt{21536}} = \frac{-6228}{11137.66} = -0.559.$$

Therefore, it is a case of moderate degree of negative correlation.

Correlation of Bivariate Grouped Data

When we have to find coefficient of correlation from a bivariate grouped data table, the following formula is applicable :

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

This formula is the same as that of (*iv*). The only difference is that here the deviations are also multiplied by the frequencies.

The following illustration shall explain the application of this formula :

Illustration 5. Find the coefficient of correlation between the age and the sum assured from the following table :

	Sum assured (in Rs.)				
Age group	10,000	20,000	30,000	40,000	50,000
20–30	4	6	3	7	1
30–40	2	8	15	7	1
40–50	3	9	12	6	2
50–60	8	4	2	—	—

(MBA, Delhi Univ., 1999)

Solution. Let the sum assured be denoted by X and the age group by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

Y		X		*10,000*	*20,000*	*30,000*	*40,000*	*50,000*				
		d_y	d_x	–2	–1	0	1	2	f	fd_y	fd_y^2	fd_xd_y
20–30	*m.p.* 25	–2		4 [16]	6 [12]	3 [0]	7 [–14]	1 [–4]	21	–42	84	+10
30–40	35	–1		2 [4]	8 [8]	15 [0]	7 [–7]	1 [–2]	33	–33	33	+3
40–50	45	0		3 [0]	9 [0]	12 [0]	6 [0]	2 [0]	32	0	0	0
50–60	55	+1		8 [–16]	4 [–4]	2 [0]	—	—	14	+14	14	–20
			f	17	27	32	20	4	$N = 100$	$\Sigma fd_y = -61$	$\Sigma fd_y^2 = 131$	$\Sigma fd_xd_y = -7$
			fd_x	–34	–27	0	20	8	$\Sigma fd_x = -33$			
			fd_x^2	68	27	0	20	16	$\Sigma fd_x^2 = 131$			
			fd_xd_y	4	16	0	–21	–6	$\Sigma fd_xd_y = -7$			

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{100(-7) - (-33)(-61)}{\sqrt{100(131) - (-33)^2}\sqrt{100(131) - (-61)^2}}$$

$$= \frac{-700-2013}{\sqrt{13100-1089}\sqrt{13100-3721}} = \frac{-2713}{\sqrt{12011}\sqrt{9379}}$$

$$= \frac{-2713}{109.59 \times 96.85} = -0.256.$$

Hence the age and sum assured are negatively correlated, *i.e.*, as age goes up the sum assured comes down.

Illustration 6. Calculate the coefficient of correlation from the following bivariate frequency distribution :

Sales Revenue	*Advertising Expenditure (Rs. '000)*			
(Rs. lakhs)	*5–10*	*10–15*	*15–20*	*20–25*
75–125	4	1	—	—
125–175	7	6	2	1
175–225	1	3	4	2
225–275	1	1	3	4

Solution. Let sales revenue be denoted by Y and advertising expenditure by X.

CALCULATION OF COEFFICIENT OF CORRELATION

Y \ X		*m.p.*	5–10 7.5	10–15 12.5	15–20 17.5	20–25 22.5				
		d_y \ d_x	−1	0	+1	+2	f	fd_y	fd_y^2	fd_xd_y
75–125	*m.p.* 100	−2	4 | 8	1 | 0	—	—	5	−10	20	8
125–175	150	−1	7 | 7	6 | 0	2 | −2	1 | −2	16	−16	16	3
175–225	200	0	1 | 0	3 | 0	4 | 0	2 | 0	10	0	0	0
225–275	250	+1	1 | −1	1 | 0	3 | 3	4 | 8	9	9	9	10
		f	13	11	9	7	N = 40	$\Sigma fd_y = -17$	$\Sigma fd_y^2 = 45$	$\Sigma fd_xd_y = 21$
		fd_x	−13	0	+9	+14	$\Sigma fd_x = 10$			
		fd_x^2	13	0	9	28	$\Sigma fd_x^2 = 50$			
		fd_xd_y	14	0	1	6	$\Sigma fd_xd_y = 21$			

$$r = \frac{N\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{40 \times 21 - 10(-17)}{\sqrt{40 \times 50 - (10)^2}\sqrt{40 \times 45 - (-17)^2}}$$

$$= \frac{840+170}{\sqrt{1900 \times 1511}} = \frac{1010}{1694.373} = 0.596$$

There is a moderate degree of positive correlation between sales revenue and advertising expenditure.

Assumptions of the Pearsonians Coefficient

The Karl Pearson's coefficient of correlation is based on the following assumptions :

1. There is linear relationship between the variables, *i.e.*, when the two variables are plotted on a scatter diagram, a straight line will be formed by the points so plotted.

2. The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc., are affected by such forces that a normal distribution is formed.

3. There is a cause-and-effect relationship between the forces affecting the distribution of the items in the two series. If such a relationship is not formed between the variables, *i.e.,* if the variables are independent there cannot be any correlation. For example, there is no relationship between income and height because the forces that affected these variables are common.

Properties of the Coefficient of Correlation

The following are the important properties of the coefficient of correlations, r :

1. The coefficient of correlation lies between –1 and +1. Symbolically, $-1 \le r \le +1$ or $|r| \le 1$.

Proof. $$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2}\ \sqrt{\Sigma (Y - \bar{Y})^2}}$$

Let $$a = \frac{(X - \bar{X})}{\sqrt{\Sigma (X - \bar{X})^2}},\ b = \frac{(Y - \bar{Y})}{\sqrt{\Sigma (Y - \bar{Y})}}$$

Then $$\Sigma (a + b)^2 = \Sigma a^2 + 2\Sigma ab + \Sigma b^2$$

$$= 1 + 2r + 1 = 2(1 + r) \ge 0 \quad \text{or} \quad 1 + r \ge 0 \qquad ...(i)$$

Similarly, $$\Sigma (a - b)^2 = \Sigma a^2 - 2\Sigma ab + \Sigma b^2$$

$$= 1 - 2r + 1 = 2(1 - r) \ge 0 \quad \text{or} \quad 1 - r \ge 0 \qquad ...(ii)$$

From (i) and (ii), $-1 \le r \le 1$.

2. The coefficient of correlation is independent of change of origin and scale.

Proof. By change of origin we mean subtracting some constant from the given value of X and Y and by change of scale we mean dividing or multiplying every value of X and Y by some constant.

We know that $$r = \frac{\Sigma (X - \bar{X})\ (Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2\ \ \Sigma (Y - \bar{Y})^2}} \qquad ...(i)$$

where $\bar{X}$ and $\bar{Y}$ refer to the acutal means of X and Y series.

Let us now change the origin and scale. Deduct a fixed quantity a from X and b from Y. Also divide X and Y series by a fixed value i and c.

Let the new values be denoted by u and v.

$$u = \frac{X - a}{i} \qquad v = \frac{Y - b}{c}$$

$$X = a + iu, \qquad Y = b + cv$$

$$\bar{X} = a + i\bar{u}, \qquad \bar{Y} = b + c\bar{v}$$

$$X - \bar{X} = i(u - \bar{u}) \qquad Y - \bar{Y} = c(v - \bar{v})$$

Substituting these values in (i), we get

$$\frac{\Sigma (u - \bar{u})\ (v - \bar{v})}{\sqrt{\Sigma (u - \bar{u})^2}\ \sqrt{\Sigma (v - \bar{v})^2}}$$

Thus the formula for r remains unchanged. Hence the value of r is independent of change of origin and scale.

3. The coefficient of correlation is the geometric mean of two regression coefficients.*

Symbolically : $r = \sqrt{b_{xy} \times b_{yx}}$

4. If X and Y are independent variables then coefficient of correlation is zero. However, the converse is not true.

Interpreting the Coefficient of Correlation

The coefficient of correlation measures the degree of relationship between two sets of figures. As the reliability of estimates depends upon the closeness of the relationship it is imperative that utmost care be taken while interpreting the value of coefficient of correlation otherwise fallacious conclusions be drawn.

Unfortunately, the interpretation of the coefficient of correlation depends very much on experience. The full significance of r will only be grasped after working out a number of correlation problems and seeing the kind of data that give rise to various values of r. The investigator must know his data thoroughly in order to avoid errors of interpretation. He must be familiar, or become familiar, with all the relationships and theory which bear upon the data and should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. However, the following general guidelines are given which would help in interpreting the value of r.

1. When $r = +1$, it means there is perfect positive correlation between the variables.

2. When $r = -1$, it means there is perfect negative correlation between the variables.

3. When $r = 0$, it means there is no correlation between the variables, *i.e.*, the variables are uncorrelated.

4. The closer r is to $+1$ or -1, the closer the relationship between the variables and the closer r is to 0, the less closer the relationship. Beyond this is not safe to go. The full interpretation of r depends upon circumstances, one of which is the size of the sample. All that can really be said that when estimating the value of one variable from the value of another; the higher the value of r, the better the estimate.

5. The closeness of the relationship is not proportional to r. If the value of r is 0.8, it does not indicate a relationship twice as close as that of 0.4. It is in fact very much closer.

Coefficient of Correlation and Probable Error

The probable error of the coefficient of correlation helps in interpreting its value. With the help of probable error it is possible to determine the liability of the value of the coefficient in so far as it depends on the condition of random sampling. The probable error of the coefficient of correlation is obtained as follows :

$$\text{P.E.}r^* = 0.6745 \frac{1 - r^2}{\sqrt{N}}$$

where r is the coefficient of correlation and N the number of pairs of items.

1. If the value of r is less than the probable error, there is no evidence of correlation, *i.e.*, the value of r is not at all significant.

---

*See chapter on Regression Analysis.

*If 0.6745 is omitted from the formula of probable error, we get the standard error from the coefficient of correlation. The standard error of r, therefore, is

$$\text{S.E.}r = \frac{1 - r^2}{\sqrt{N}}$$

2. If the value of r is more than six times the probable error, the existence of correlation is practically certain, *i.e.*, the value of r is significant.

3. By adding and subtracting the value of probable error from the coefficient of correlation we get respectively the upper and lower limits within which coefficient of correlation in the population can be expected to lie. Symbolically,

$$\rho = r \pm \text{P.E.}r$$

where ρ (rho) denotes correlation in the population.

Carrying out the computation of the probable error, assuming a coefficient of correlation of 0.80 computed from a sample of 16 pairs of items, we have

$$\text{P.E.}r = 0.6745 \frac{1-(0.8)^2}{\sqrt{16}} = 0.06$$

The limits of the correlation in the population should be $r \pm \text{P.E.} = 0.8 + 0.06 = 0.74 - 0.86$.

Instances are quite common wherein a correlation coefficient of 0.5 or even 0.4 is obviously considered to be a fairly high degree of correlation by a research worker. Yet a correlation coefficient of 0.5 means that only 25 per cent of the variation is explained. A correlation coefficient of 0.4 means that only 16 per cent of the variation is explained.

Conditions for the Use of Probable Error

The measure of probable error can be properly used only when the following three conditions exist :

1. The data must approximate to a normal frequency curve (bell-shaped curve).

2. The statistical measure for which the P.E. is computed must have been calculated from a sample.

3. The sample must have been selected in an unbiased manner and the individual items must be independent.

However, these conditions are generally not satisfied and as such the reliability of the correlation coefficient is determined largely on the basis of exterior tests of reasonableness which are often of a statistical character.

Illustration 7. If $r = 0.6$ and $N = 64$, find out the probable error of the coefficient of correlation and determine the limits for r.

Solution :

$$\text{P.E.}r = 0.6745 \frac{1-r^2}{\sqrt{N}}; \quad r = 0.6 \text{ and } N = 64$$

$$\text{P.E.}r = 0.6745 \frac{1-(0.6)^2}{\sqrt{64}} = \frac{0.6745 \times 0.64}{8} = 0.054$$

limits of $r = 0.6 \pm 0.054$ or $= 0.546$ to 0.654.

Merits and Limitations of the Pearsonian Coefficient

Amongst the mathematical methods used for measuring the degree of relationship, Karl Pearson's method is most popular. The correlation coefficient summarizes in one figure not only the degree of correlation but also the degree, *i.e.*, whether correlation is positive or negative.

However, the utility of the coefficient depends in part on a wide knowledge of the meaning of this 'yardstick' together with its limitations. The chief *limitations* of the method are :

1. The correlation coefficient always assumes linear relationship regardless of the fact whether that assumption is true or not.

2. Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.

3. The value of the coefficient is unduly affected by the extreme values.

4. As compared to other methods of finding correlation, this method is more time-consuming.

Coefficient of Determination*

One very convenient and useful way of interpreting the value of coefficient of correlation between two variables is to use the square of coefficient of correlation, which is called coefficient of determination. The coefficient of determination thus equals r^2. The coefficient, r^2 expresses the proportion of the variance in Y determined in X, that is, the ratio of the explained variance to the total variance.

Therefore, the coefficient of determination expresses the proportion of the total variation that has been "explained", or the relative reduction in variance when measured about the regression equation rather than about the mean of the dependent variable. If the value of $r = 0.9$, r^2 will be 0.81 and this would mean that 81 per cent of the variation in the dependent variable has been explained by the independent variable. The maximum value of r^2 is unity because it is possible to explain all of the variation in Y, but it is not possible to explain more than all of it.

It is much easier to understand the meaning of r^2 and r and, therefore, the coefficient of determination is to be preferred in presenting the result of correlation analysis. Tuttle has beautifully pointed out that "the coefficient of correlation has been grossly overrated and is used entirely too much. Its square coefficient of determination is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between the two correlated variables."

The relationship between r and r^2 may be noted—as the value of r decreases from its maximum value of l, the value of r^2 decreases much more rapidly. r will of course always be larger than r^2, unless $r^2 = 0$ or 1, when $r = r^2$,

r	r^2	r	r^2
0.90	0.81	0.60	0.36
0.80	0.64	0.50	0.25
0.70	0.49	0.40	0.16

Thus the coefficient of correlation is 0.707 when just half the variance in Y is due to X.

It should be clearly noted that the fact that a correlation between two variables has a value of $r = 0.60$ and the correlation between two other variables has a value of $r = 0.30$ does not demonstrate that the first correlation is twice as strong as the second. The relationship between the two given values of r can better be understood by computing the value of r^2. When $r = 0.6$, $r^2 = 0.36$ and when $r = 0.30$, $r^2 = 0.09$.

The coefficient of determination is a highly useful measure. However, it is often misinterpreted. The term itself may be misleading in that it implies that the variable X stands in a determining of casual relationship of the variable Y. The statistical evidence itself never establishes the existence of such causality. All that statistical evidence can do it to define covariation, that term being used in a perfectly

*$1 - r^2$ is known as the coefficient of non-determination.

neutral sense. Whether causality is present or not and which way it runs if it is present, must be determined on the basis of evidence other than the quantitative observations.

III. RANK CORRELATION COEFFICIENT

This method of finding out covariability or the lack of it between two variables was developed by the British psychologist Charles, Edward Spearman in 1904. This measure is especially useful when quantitative measure of certain factors (such as in the evaluation of leadership ability or the judgement of female beauty) cannot be fixed, but the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank* in the group. In any event, the rank correlation coefficient is applied to a set of ordinal rank numbers, with 1 for the individual ranked first in quantity or quality, and so on, N for the individual ranked last in a group of N individuals (or N pairs of individuals). Spearman's rank correlation coefficient is defined as :

$$R = 1 - \frac{6\Sigma D^2}{N\left(N^2 - 1\right)} \text{ or } 1 - \frac{6\Sigma D^2}{\left(N^3 - N\right)}$$

where R denotes rank coefficient of correlation and D refers to the difference of ranks between paired items in two series.

The value of this coefficient also lies between +1 and –1. When R is +1, there is complete agreement in the order of the ranks and the ranks are in the same direction. When R is –1, there is complete agreement in the order of the ranks and they are in opposite directions. This shall be clear from the following :

R_1	R_2	D $(R_1 - R_2)$	D^2	R_1	R_2	D $(R_1 - R_2)$	D^2
1	1	0	0	1	3	–2	4
2	2	0	0	2	2	0	0
3	3	0	0	3	1	2	4
			$\Sigma D^2 = 0$				$\Sigma D^2 = 8$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 0}{3^3 - 3} = 1 - 0 = 1 \qquad R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 8}{3^3 - 3} = 1 - 2 = -1$$

In rank correlation we may have two types of problems :

A. Where actual ranks are given.

B. Where ranks are not given.

A. Where Actual Ranks are Given.

Where actual ranks are given, the steps required for computing rank correlation are :

(*i*) Take the differences of the two ranks, *i.e.*, $(R_1 - R_2)$ and denote these differences by D.

(*ii*) Square these differences and obtain the total ΣD^2.

(*iii*) Apply the formula :

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

*The rank transformation for a sample of n observation replaces the smallest observation by the integer 1 (called) the rank, the next by rank 2 and so on until the largest observation is replaced by rank n.

Illustration 7. Two managers are asked to rank a group of employees in order of potential for eventually becoming top managers. The rankings are as follows :

Employees	*Ranking by Manager I*	*Ranking by Manager II*
A	10	9
B	2	4
C	1	2
D	4	3
E	3	1
F	6	5
G	5	6
H	8	8
I	7	7
J	9	10

Compute the coefficient of rank correlation and comment on the value.

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

Employees	*Rank by Manager I* R_1	*Rank by Manager II* R_2	$(R_1 - R_2)^2$ D^2
A	10	9	1
B	2	4	4
C	1	2	1
D	4	3	1
E	3	1	4
F	6	5	1
G	5	6	1
H	8	8	0
I	7	7	0
J	9	10	1
$N = 10$			$\Sigma D^2 = 14$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 14}{990} = 1 - 0.085 = 0.915$$

Thus we find that there is a high degree of positive correlation in the ranks assigned by the two managers.

Illustration 8. Two housewives, Geeta and Rita, asked to express their preference for different kinds of detergents, gave the following replies :

Detergent	*Geeta*	*Rita*
A	4	4
B	2	1
C	1	2
D	3	3
E	7	8
F	8	7
G	6	5
H	5	6
I	9	9
J	10	10

To what extent the preferences of these two ladies go together?

Solution. In order to find out how far the preferences for different kinds of detergents go together, we will calculate rank correlation coefficient.

CALCULATION OF RANK CORRELATION COEFFICIENT

Detergent	Rank by Geeta R_1	Rank by Rita R_2	$(R_1 - R_2)^2$ D^2
A	4	4	0
B	2	1	1
C	1	2	1
D	3	3	0
E	7	8	1
F	8	7	1
G	6	5	1
H	5	6	1
I	9	9	0
J	10	10	0
$N = 10$			$\Sigma D^2 = 6$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 6}{990} = 1 - 0.036 = 0.964.$$

Thus the preferences of these two ladies agree very closely as far as their opinion on detergents is concerned.

B. Where Ranks are not Given.

When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value, we must follow the same method in case of all the variables.

Illustration 9. Calculate the rank correlation coefficient for the following data of marks of 2 tests given to candidates for a clerical job.

Preliminary test :	92	89	87	86	83	77	71	63	53	50
Final test :	86	83	91	77	68	85	52	82	37	57

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

Preliminary test X	R_1	Final test Y	R_2	$(R_1 - R_2)^2$ D^2
92	10	86	9	1
89	9	83	7	4
87	8	91	10	4
86	7	77	5	4
83	6	68	4	4
77	5	85	8	9
71	4	52	2	4
63	3	82	6	9
53	2	37	1	1
50	1	57	3	4
$N = 10$				$\Sigma D^2 = 44$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 44}{990} = 1 - 0.267 = 0.733.$$

Thus, there is a high degree of positive correlation between preliminary and final test.

Equal Ranks or Tie in Ranks

In some cases it may be found necessary to assign equal rank to two or more individuals or entries. In such a case, it is customary to give each individual or entry an average rank. Thus if two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$, that is 5.5 while if

three are ranked equal at fifth place, they are given the rank $\frac{5+6+7}{3}=6$. In other words, where two or more individuals are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have got had they differed slightly from each other.

Where equal ranks are assigned to some entries, an adjustment in the above formula for calculating the rank coefficient of correlation is made.

The adjustment consists of adding $\frac{1}{12}(m^3-m)$ to the value of ΣD^2, where m stands for the number of items whose ranks are common. If there are more than one such group of items with common rank, this value is added as many times as the number of such groups. The formula can thus be written as :

$$R=1-\frac{6\{\Sigma D^2+\frac{1}{12}(m_1^3-m_1)+\frac{1}{12}(m_2^3-m_2)+.....\}}{N^3-N}$$

Illustration 10. An examination of eight applicants for a clerical post was taken by a firm. From the marks obtained by the applicants in the Accountancy and Statistics papers, compute rank coefficient of correlation.

Applicant	:	A	B	C	D	E	F	G	H
Marks in Accountancy	:	15	20	28	12	40	60	20	80
Marks in Statistics	:	40	30	50	30	20	10	30	60

(MBA, Delhi Univ., 2009)

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

Applicants	*Marks in Accountancy* X	*Rank assigned* R_1	*Marks in Statistics* Y	*Rank assigned* R_2	$(R_1-R_2)^2$ D^2
A	15	2	40	6	16.00
B	20	3.5	30	4	0.25
C	28	5	50	7	4.00
D	12	1	30	4	9.00
E	40	6	20	2	16.00
F	60	7	10	1	36.00
G	20	3.5	30	4	0.25
H	80	8	60	8	0.00
$N=8$					$\Sigma D^2=81.5$

$$R=1-\frac{6\{\Sigma D^2+\frac{1}{12}(m_1^3-m_1)+\frac{1}{12}(m_2^3-m_2)\}}{N^3-N}$$

The item 20 is repeated 2 times in series X and hence $m_1=2$. In series Y, the item 30 occurs 3 times and hence $m_2=3$. Substituting these values in the above formula :

$$R=1-\frac{6\left\{81.5+\frac{1}{12}(2^3-2)+\frac{1}{12}(3^3-3)\right\}}{8^3-8}$$

$$=1-\frac{6(81.5+0.5+2)}{504}=1-\frac{6\times 84}{504}=0$$

There is no correlation between the marks obtained in the two subjects.

Illustration 11. Ten competitors in a beauty contest are ranked by three judges in the following order :

1st Judge	:	1	6	5	10	3	2	4	9	7	8
2nd Judge	:	3	5	8	4	7	10	2	1	6	9
3rd Judge	:	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty. *(MBA, M.D. Univ., 1996; MBA, Bharthidasan Univ., 2001)*

Solution : In order to find out which pair of judges has the nearest approach to common tastes in beauty, we compare rank correlation between the judgement of

(*i*) 1st judge and 2nd judge.
(*ii*) 2nd judge and 3rd judge.
(*iii*) 1st judge and 3rd judge.

Rank by 1st Judge R_1	*Rank by 2nd Judge* R_2	*Rank by 3rd Judge* R_3	$(R_1 - R_2)^2$ D^2	$(R_2 - R_3)^2$ D^2	$(R_1 - R_3)^2$ D^2
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
$N = 10$	$N = 10$	$N = 10$	$\Sigma D^2 = 200$	$\Sigma D^2 = 214$	$\Sigma D^2 = 60$

$$R(\text{I\&II}) = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 200}{10^3 - 10} = 1 - \frac{1200}{990} = -0.212$$

$$R(\text{II\&III}) = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 214}{10^3 - 10} = 1 - \frac{1284}{990} = -0.297$$

$$R(\text{I\&III}) = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = 0.636$$

Since coefficient of correlation is maximum in the judgment of the first and third judges, we conclude that they have the nearest approach to common tastes in beauty.

Merits and Limitations of the Rank Method

Merits. 1. This method is simpler to understand and easier to apply compared to the Karl Pearson's method. The answers obtained by this method and the Karl Pearson's method will be the same provided no value is repeated, *i.e.*, all the items are different.

2. Where the data are of a qualitative nature like honesty, efficiency, intelligence, etc., this method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and the degree of correlation established by applying the method.

3. This is the only method that can be used where we are given the ranks and not the actual data.

4. Even where actual data are given, rank method can be applied for ascertaining rough degree of correlation.

Limitations. 1. This method cannot be used for finding out correlation in a grouped frequency distribution.

2. Where the number of observations exceed 30, the calculations become quite tedious and require a lot of time. Therefore, this method should not be applied where N is exceeding 30 unless we are given the ranks and not the actual values of the variable.

When to Use Rank Correlation Coefficient

The rank method has two principal uses :

(1) The initial data are in the form of ranks.

(2) If N is fairly small (say, not greater than 25 or 30) rank method is sometimes applied to interval data as an approximation to the more time-consuming r. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30, the labour required in ranking the scores becomes greater than what is justified by the anticipated saving of time through the rank formula.

Illustration 12. The coefficient of rank correlation between debenture prices and share prices is found to be 0.143. If the sum of squares of the differences in rank is given to be 48, find the value of N.

Solution. $$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

where $$R = 0.143,\ \Sigma D^2 = 48$$

$$0.143 = 1 - \frac{6 \times 48}{N^3 - N}$$

$$\frac{288}{N^3 - N} = 0.857 \text{ or } 0.857\,(N^3 - N) = 288$$

$$(N^3 - N) = \frac{288}{0.857} = 336$$

or $$N^3 - N - 336 = 0 \text{ or } N^3 - N - 343 + 7 = 0$$

$$(N-7)(N^2 + 7N) + 48(N-7) = 0$$

or $$(N-7)(N^2 + 7N + 48) = 0$$

either $$N - 7 = 0 \quad i.e.,\ N = 7 \quad \text{or} \quad N^2 + 7N + 48 = 0$$

Since $b^2 - 4ac$ is negative, value of N belongs to the set of complex numbers. Hence $N = 7$.

Illustration 13. The coefficient of rank correlation of the marks obtained by 10 students in statistics and accountancy was found to be 0.8. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 7 instead of 9. Find the correct coefficient of rank correlation.

Solution. $$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$0.8 = 1 - \frac{6\Sigma D^2}{10^3 - 10} \text{ or } 0.8 = 1 - \frac{6\Sigma D^2}{990} \text{ or } \frac{6\Sigma D^2}{990} = 0.2$$

$$6\Sigma D^2 = 198 \quad \text{or} \quad \Sigma D^2 = 33$$

But this is not correct ΣD^2

$$\text{Correct } \Sigma D^2 = 33 - (7)^2 + (9)^2 = 65$$

$$R = 1 - \frac{6 \times 65}{990} = 1 - \frac{390}{990} = 1 - 0.394 = 0.606$$

Thus the correct value of the rank correlation coefficient is 0.606.

IV. METHOD OF LEAST SQUARES

For finding out correlation by the coefficient method of least squares we have to calculate the values of two regression coefficients—that of x on y and y on x. The correlation coefficient is the square root of the product of two regression coefficients. Symbolically,

$$r^* = \sqrt{b_{xy} \times b_{yx}}$$

Lag and Lead in Correlation

The study of lag and lead is of special significance while studying economic and business series. In the correlation of time series the investigator may find that there is a time gap before a cause-and-effect

*For details of this method refer to next chapter on Regression Analysis.

relationship is established. For example, the supply of a commodity may increase today, but it may not have an immediate effect on prices—it may take a few days or even months for prices to adjust to the increased supply. The difference in the period before a cause-and-effect relationship is established is called 'Lag'. While computing correlation this time gap must be considered; otherwise, fallacious conclusions may be drawn. The pairing of items is adjusted according to the time lag.

If the supply affects the prices, say, after 5 months, then the pairing would be done as follows :

Months	*Supply*	*Price*
Jan.	100	70
Feb.	105	69
March	108	80
April	112	72
May	118	75
June	120	70
July	125	74
Aug.	104	75
Sept.	112	78
Oct.	116	80
Nov.	122	78
Dec.	127	75

Taking the new pairs of values, correlation can be calculated in the same manner as discussed earlier.

Illustration 14. The following are the monthly figures of advertising expenditure and sales of a firm. It is generally found that advertising expenditure has its impact on sales generally after two months. Allowing for this time lag, calculate coefficient of correlation between expenditure on advertisement and sales.

Month	*Advertising expenditure*	*Sales (Rs.)*	*Month*	*Advertising expenditure*	*Sales (Rs.)*
Jan.	50	1,200	July	140	2,400
Feb.	60	1,500	Aug.	160	2,600
March	70	1,600	Sept.	170	2,800
April	90	2,000	Oct.	190	2,900
May	120	2,200	Nov.	200	3,100
June	150	2,500	Dec.	250	3,900

Solution. Allow for a time lag of 2 months, *i.e.*, link advertising expenditure of January with sales for March, and so on.

CALCULATION OF CORRELATION COEFFICIENT

Month	*Advertising expenditure* X	$(X-\bar{X})/10$ x	x^2	*Sales* Y	$(Y-\bar{Y})/100$ y	y^2	xy
Jan.	50	–7	49	1,600	–10	100	70
Feb.	60	–6	36	2,000	–6	36	36
March	70	–5	25	2,200	–4	16	20
April	90	–3	9	2,500	–1	1	3
May	120	0	0	2,400	–2	4	0
June	150	+3	9	2,600	0	0	0
July	140	+2	4	2,800	+2	4	4
Aug.	160	+4	16	2,900	+3	9	12
Sept.	170	+5	25	3,100	+5	25	25
Oct.	190	+7	49	3,900	+13	169	91
	$\Sigma X = 1,200$	$\Sigma x = 0$	$\Sigma x^2 = 222$	$\Sigma Y = 26,000$	$\Sigma y = 0$	$\Sigma y^2 = 364$	$\Sigma xy = 261$

$$\bar{X} = \frac{1,200}{10} = 120, \quad \bar{Y} = \frac{26,000}{10} = 2,600$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{261}{\sqrt{222 \times 364}} = \frac{261}{284.27} = 0.918$$

There is a very high degree of positive correlation between advertising expenditure and sales.

MISCELLANEOUS ILLUSTRATIONS

Illustration 15. A Computer while calculating the correlation coefficient between the variables X and Y obtained following results :

$$N = 30,\ \Sigma X = 120,\ \Sigma X^2 = 600,\ \Sigma Y = 90,\ \Sigma Y^2 = 250,\ \Sigma XY = 335$$

It was, however, later discovered at the time of checking that it had copied down two pairs of observations as:

X	Y
8	10
12	7

While the correct values were

X	Y
8	12
10	8

Obtain the correct value of the correlation coefficient between X and Y.
(MBA, Vikram, Univ.; MBA, Kumaun Univ., 2007)

Solution.

Correct $\Sigma X = 120 - 8 - 12 + 8 + 10 = 120 - 2 = 118$

Correct $\Sigma Y = 90 - 10 - 7 + 12 + 8 = 93$

Correct $\Sigma X^2 = 600 - (8)^2 - (12)^2 + (8)^2 + (10)^2 = 600 - 64 - 144 + 64 + 100 = 556$

Correct $\Sigma Y^2 = 250 - (10)^2 - (7)^2 + (12)^2 + (8)^2 = 250 - 100 - 49 + 144 + 64 = 309$

Correct $\Sigma XY = 335 - (8 \times 10) - (12 \times 7) + (8 \times 12) + (10 \times 8)$

$= 335 - 80 - 84 + 96 + 80 = 347$

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{(30 \times 347) - (118 \times 93)}{\sqrt{(30 \times 556) - (118)^2}\sqrt{30 \times 309 - (93)^2}}$$

$$= \frac{10410 - 10974}{\sqrt{16680 - 13924}\sqrt{9270 - 8649}} = \frac{-564}{\sqrt{2756}\sqrt{621}} = \frac{-564}{52.50 \times 24.92} = -0.43$$

Thus the correct value of correlation coefficient between X and Y is -0.43.

Illustration 16. Coefficient of correlation between two variates X and Y is 0.3. Their covariance is 9. The variance of X is 16. Find the standard deviation of Y series.

Solution. Covariance is given by $\frac{\Sigma xy}{N}$, where x and y are the deviations of X and Y series from their respective means.

Variance of X series is 16, or $\sigma_x = \sqrt{16} = 4$

Substituting the given value in the formula $r = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$, we get

$0.3 = 9 \times \frac{1}{4\sigma_y}$ or $1.2\sigma_y = 9$. Hence $\sigma_y = 9/1.2 = 7.5$.

Illustration 17. Family income and its percentage spent on food in the case of one hundred families gave the following bivariate frequency distribution. Calculate the coefficient of correlation and interpret its value.

Food Expenditure	*Monthly Family Income (Rs. '000's)*				
(in%)	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30
10–15	—	—	—	3	7
15–20	—	4	9	4	3
20–25	7	6	12	5	—
25–30	3	10	19	8	—

(MBA, Delhi Univ., 2003)

Solution. Let family income be denoted by X and food expenditure (in %) by Y.

CALCULATION OF CORRELATION COEFFICIENT

Y \ X		m.p.	5–10 7.5	10–15 12.5	15–20 17.5	20–25 22.5	25–30 27.5				
		d_x / d_y	–2	–1	0	1	2	f	fd_y	fd_y^2	fd_xd_y
10–15	m.p. 12.5	–1				3 (–3)	7 (–14)	10	–10	10	–17
15–20	17.5	0		4 (0)	9 (0)	4 (0)	3 (0)	20	0	0	0
20–25	22.5	1	7 (–14)	6 (–6)	12 (0)	5 (5)		30	30	30	–15
25–30	27.5	2	3 (–12)	10 (–20)	19 (0)	8 (16)		40	80	160	–16
		f	10	20	40	20	10	$N = 100$	$\Sigma fd_y = 100$	$\Sigma fd_y^2 = 200$	$\Sigma fd_xd_y = -48$
		fd_x	–20	–20	0	20	20	$\Sigma fd_x = 0$			
		fd_x^2	40	20	0	20	40	$\Sigma fd_x^2 = 120$			
		fd_xd_y	–26	–26	0	18	–14	$\Sigma fd_xd_y = -48$			

$$r = \frac{N\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{(100 \times -48) - (0 \times 100)}{\sqrt{100 \times 120 - (0)^2}\sqrt{100 \times 200 - (100)^2}}$$

$$= \frac{-4800}{\sqrt{12000}\sqrt{10000}} = \frac{-48}{\sqrt{120}\sqrt{100}} = -0.438.$$

There seems to be a low degree of negative correlation between family income and its percentage spent on food expenditure.

Illustration 18. An office contains 12 clerks. The long serving clerks feel that they should have a seniority increment based on length of service built into their salary structure. An assessment of their efficiency by their departmental manager and the personnel department produces a ranking of efficiency. This is shown below together with a ranking of their length of service. Do the data support the clerks, claim for seniority increment ?

Ranking according to length of service	:	1	2	3	4	5	6	7	8	9	10	11	12
Ranking according to efficiency	:	2	3	5	1	9	10	11	12	8	7	6	4

Solution. CALCULATION OF RANK CORRELATION

Ranking according to length of service R_1	*Ranking according to efficiency* R_2	$(R_1 - R_2)$ D	D^2
1	2	−1	1
2	3	−1	1
3	5	−2	4
4	1	+3	9
5	9	−4	16
6	10	−4	16
7	11	−4	16
8	12	−4	16
9	8	+1	1
10	7	+3	9
11	6	+5	25
12	4	+8	64
$N = 12$			$\Sigma D^2 = 178$

Rank correlation coefficient is given by :

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 178}{12^3 - 12} = 1 - \frac{1068}{1716} = 1 - 0.622 = 0.378$$

Since there is a low degree of positive correlation between length of service and efficiency, the clerks' claim does not justify for a seniority increment based on the length of service.

Illustration 19. The ranks of the same 15 students in two subjects *A* and *B* are given below, the two numbers within the brackets denoting the ranks of the same student in *A* and *B* respectively :

(1, 10), (2, 7), (3, 2), (4, 6), (5, 4), (6, 8), (7, 3), (8, 1), (9, 11), (10, 15), (11, 9), (12, 5), (13, 14), (14, 12), (15, 13)

Use Spearman's formula to find the rank correlation coefficient. *(MBA, Sukhadia Univ., 2008)*

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

Rank of A R_1	*Rank of B* R_2	$(R_1 - R_2)^2$ D^2
1	10	81
2	7	25
3	2	1
4	6	4
5	4	1
6	8	4
7	3	16
8	1	49
9	11	4
10	15	25
11	9	4
12	5	49
13	14	1
14	12	4
15	13	4
		$\Sigma D^2 = 272$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 272}{15^3 - 15} = 1 - .486 = 0.514.$$

There is a moderate degree of positive correlation between the ranks in subject *A* and *B*.

Illustration 20. Calculate the coefficient of correlation from the following data :

Adv. Exp. (Rs. Lakhs)	X :	10	12	15	23	20
Sales (Rs. Crores)	Y :	14	17	23	25	21

Solution. CALCULATION OF COEFFICIENT OF CORRELATION

Adv. Exp. X	$(X-16)$ x	x^2	*Sales* Y	$(Y-20)$ y	y^2	xy
10	−6	36	14	−6	36	+36
12	−4	16	17	−3	9	+12
15	−1	1	23	+3	9	−3
23	+7	49	25	+5	25	+35
20	+4	16	21	+1	1	+4
$\Sigma X = 80$	$\Sigma x = 0$	$\Sigma x^2 = 118$	$\Sigma Y = 100$	$\Sigma y = 0$	$\Sigma y^2 = 80$	$\Sigma xy = 84$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{84}{\sqrt{118 \times 80}} = \frac{84}{97.16} = +0.865.$$

There is a high degree of positive correlation between sales and advt. expenditure.

Illustration 21. Calculate coefficient of correlation from the following data taking deviation from 48 in case of X series and 20 in case of Y series :

X :	40	42	46	48	50	56
Y :	10	12	15	23	27	30

Solution. CALCULATION OF CORRELATION COEFFICIENT

X	$(X-48)$ d_x	d_x^2	Y	$(Y-20)$ d_y	d_y^2	$d_x d_y$
40	−8	64	10	−10	100	+80
42	−6	36	12	−8	64	+48
46	−2	4	15	−5	25	+10
48	0	0	23	+3	9	0
50	+2	4	27	+7	49	+14
56	+8	64	30	+10	100	+80
	$\Sigma d_x = -6$	$\Sigma d_x^2 = 172$		$\Sigma d_y = -3$	$\Sigma d_y^2 = 347$	$\Sigma d_x d_y = 232$

$$= \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$r = \frac{6 \times 232 - (-6)(-3)}{\sqrt{6(172) - (-6)^2}\sqrt{6(347) - (-3)^2}}$$

$$= \frac{1392 - 18}{\sqrt{1032 - 36}\sqrt{2082 - 9}} = \frac{1374}{\sqrt{996}\sqrt{2073}} = \frac{1374}{31.56 \times 45.53} = 0.956$$

Illustration 22. A panel of men and a panel of women were asked by a consumer testing organisation to rank 8 brands of tea according to taste. A rank of 1 was given to the best tasting tea and a rank of 8 to the worst.

Brand	:	A	B	C	D	E	F	G	H
Panel of Women (X)	:	5	4	3	6	7	8	1	2
Panel of Men (Y)	:	4	5	6	3	8	7	2	1

Determine how closely men's and women's tastes in tea are related.

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

R_1	R_2	$(R_1 - R_2)^2$
5	4	1
4	5	1
3	6	9
6	3	9
7	8	1
8	7	1
1	2	1
2	1	1
		$\Sigma D^2 = 24$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 24}{8^3 - 8} = 1 - \frac{144}{512 - 8} = 1 - 0.286 = 0.714$$

There is a high degree of positive correlation between the tastes of men and women in tea.

Illustration 23. A company gives on-the-job training to its salesmen which is followed by a test. It is considering whether it should terminate the services of any salesman who does not do well in the test.

The following data give the test scores and sales made by nine salesmen during the last one year :

Test scores	:	14	19	24	21	26	22	15	20	19
Sales (Rs. '000):		31	36	48	37	50	45	33	41	39

Compute the coefficient of correlation between test scores and sales. Does it indicate that termination of the services of salesman with low test scores is justified? ***(MBA, Madurai-Kamaraj Univ., 2007)***

Solution. CALCULATION OF CORRELATION COEFFICIENT

Test scores X	$(X - 20)$ x	x^2	*Sales* Y	$(Y - 40)$ y	y^2	xy
14	–6	36	31	–9	81	+54
19	–1	1	36	–4	16	+4
24	+4	16	48	+8	64	+32
21	+1	1	37	–3	9	–3
26	+6	36	50	+10	100	+60
22	+2	4	45	+5	25	+10
15	–5	25	33	–7	49	+35
20	0	0	41	+1	1	0
19	–1	1	39	–1	1	+1
$\Sigma X = 180$	$\Sigma x = 0$	$\Sigma x^2 = 120$	$\Sigma Y = 360$	$\Sigma y = 0$	$\Sigma y^2 = 346$	$\Sigma xy = 193$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}, \text{ where } x = (X - \bar{X}); y = (Y - \bar{Y})$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{180}{9} = 20; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{360}{9} = 40$$

Since the actual means of X and Y are whole numbers, we should take deviation from actual means of X and Y to simplify the calculations.

Substituting the values

$$r = \frac{193}{\sqrt{120 \times 346}} = \frac{193}{203.76} = 0.947$$

There is a high degree of positive correlation between test scores and sales. It does not indicate that the termination of the services of salesman with low test scores is justified.

Illustration 24. Find the correlation coefficient between age and playing habits of the following students :

Age	:	15	16	17	18	19	20
No. of students	:	250	200	150	120	100	80
Regular players	:	200	150	90	48	30	12

Solution. Let us find the percentage of regular players and then calculate coefficient of correlation between age and percentage.

X	$(X-17)$ d_x	d_x^2	No. of Students	Regular Players Y	% of Regular Players	$(Y-50)$ d_y	d_y^2	d_xd_y
15	–2	4	250	200	80	+30	900	–60
16	–1	1	200	150	75	+25	625	–25
17	0	0	150	90	60	+10	100	0
18	+1	1	120	48	40	–10	100	–10
19	+2	4	100	30	30	–20	400	–40
20	+3	9	80	12	15	–35	1225	–105
	$\Sigma d_x = +3$	$\Sigma d_x^2 = 19$				$\Sigma d_y = 0$	$\Sigma d_y^2 = 3350$	$\Sigma d_xd_y = -240$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{6(-240) - (3)(0)}{\sqrt{6\times 19 - (3)^2}\sqrt{6\times 3350 - (0)^2}} = \frac{-1440}{\sqrt{105\times 20100}} = \frac{-1440}{1452.76} = -0.991$$

Thus there is a high degree of negative correlation between age and playing habits.

Illustration 25. Calculate Karl Pearson's coefficient of correlation from the following data and interpret its value :

Roll No.	:	1	2	3	4	5
Marks in Accountancy	:	48	35	17	23	47
Marks in Statistics	:	45	20	40	25	45

Solution. Let marks in accountancy be denoted by X and that in statistics by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

X	$(X-34)$ x	x^2	Y	$(Y-35)$ y	y^2	xy
48	+14	196	45	+10	100	+140
35	+1	1	20	–15	225	–15
17	–17	289	40	+5	25	–85
23	–11	121	25	–10	100	+110
47	+13	169	45	+10	100	+130
$\Sigma X = 170$	$\Sigma x = 0$	$\Sigma x^2 = 776$	$\Sigma Y = 175$	$\Sigma y = 0$	$\Sigma y^2 = 550$	$\Sigma xy = 280$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{280}{\sqrt{776\times 550}} = \frac{280}{653.3} = +0.429$$

It is a moderate case of positive correlation between marks in accountancy and statistics.

Illustration 26. Calculate the coefficient of correlation and its probable error from the following :

S.No.	Subject	% marks in final year exams.	% marks in sessionals
1	Hindi	75	62
2	English	81	68
3	Physics	70	65

4	Chemistry	76	60	
5	Maths.	77	69	
6	Statistics	81	72	
7	Botany	84	76	
8	Zoology	75	72	*(MBA, Jodhpur Univ., 2001)*

Solution. Let % marks in final year exams. be denoted by X and % marks in sessionals by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

X	$(X-77)$ d_x	d_x^2	Y	$(Y-68)$ d_y	d_y^2	d_xd_y
75	−2	4	62	−6	36	+12
81	+4	16	68	0	0	0
70	−7	49	65	−3	9	+21
76	−1	1	60	−8	64	+8
77	0	0	69	+1	1	0
81	+4	16	72	+4	16	+16
84	+7	49	76	+8	64	+56
75	−2	4	72	+4	16	−8
$\Sigma X = 619$	$\Sigma d_x = 3$	$\Sigma d_x^2 = 139$	$\Sigma Y = 544$	$\Sigma d_y = 0$	$\Sigma d_y^2 = 206$	$\Sigma d_xd_y = 105$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{8\times 105 - (3\times 0)}{\sqrt{8\times 139 - (3)^2}\sqrt{8\times 206}} = \frac{840}{\sqrt{1103\times 1648}} = \frac{840}{1348.237} = 0.623$$

Probable error is given by :

$$PEr = .6745\,\frac{1-r^2}{\sqrt{N}} = .6745\,\frac{1-(.623)^2}{\sqrt{8}} = \frac{.6745\times .6119}{2.8284} = 0.146$$

Illustration 27. Following figures give the rainfall in inches for the year and the production in 00's of kgs. for the Rabi crop and Kharif crop. Calculate the Karl Pearson's coefficient of correlation between rainfall and total production :

Rainfall	:	20	22	24	26	28	30	32
Rabi Production	:	15	18	20	32	40	39	40
Kharif Production	:	15	17	20	18	20	21	15

Solution. Let rainfall be denoted by X and production by Y.

CALCULATION OF CORRELATION COEFFICIENT

X	$(X-26)$ d_x	d_x^2	Y	$(Y-47)$ d_y	d_y^2	d_xd_y
20	−6	36	30	−17	289	+102
22	−4	16	35	−12	144	+48
24	−2	4	40	−7	49	+14
26	0	0	50	+3	9	0
28	+2	4	60	+13	169	+26
30	+4	16	60	+13	169	+52
32	+6	36	55	+8	64	+48
$\Sigma X = 182$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 112$	$\Sigma Y = 330$	$\Sigma d_y = 1$	$\Sigma d_y^2 = 893$	$\Sigma d_xd_y = 290$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{(7)(290) - (0)(1)}{\sqrt{(7)(112) - (0)^2}\sqrt{(7)(893) - (1)^2}} = \frac{2030}{\sqrt{784}\sqrt{6250}} = \frac{2030}{2213.594} = 0.917$$

It is a case of very high degree of positive correlation between rainfall and Agricultural production.

Illustration 28. Calculate the coefficient of correlation between weight and height for the following bivariate frequency distribution :

Weight (pounds)	*Height (inches)* 40–44	44–48	48–52	52–56	56–60	*Total*
35–55	4	40	60	—	—	104
55–75	—	—	24	88	12	124
75–95	—	—	—	8	40	48
95–115	—	—	—	—	12	12
115–135	—	—	—	4	—	4
135–155	—	—	—	4	4	8
Total	4	40	84	104	68	300

Solution. CALCULATION OF CORRELATION COEFFICIENT

Y \ X	d_y \ d_x	40–44	44–48	48–52	52–56	56–60	f	fd_y	fd_y^2	fd_xd_y
		–2	–1	0	+1	+2				
35–55	–2	16 4	80 40	0 60	—	—	104	–208	416	96
55–75	–1	—	—	0 24	–88 88	–24 12	124	–124	124	–112
75–95	0	—	—	—	0 8	0 40	48	0	0	0
95–115	+1	—	—	—	—	24 12	12	12	12	24
115–135	+2	—	—	—	8 4	—	4	8	16	8
135–155	+3	—	—	—	12 4	24 4	8	24	72	36
	f	4	40	84	104	68	$N = 300$	$\Sigma fd_y = -288$	$\Sigma fd_y^2 = 640$	$\Sigma fd_xd_y = 52$
	fd_x	–8	–40	0	104	136	$\Sigma fd_x = 192$			
	fd_x^2	16	40	0	104	272	$\Sigma fd_x^2 = 432$			
	fd_xd_y	16	80	0	–68	24	$\Sigma fd_xd_y = 52$			

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{(300) \times (52) - (192)(-288)}{\sqrt{(300)(432) - (192)^2}\sqrt{(300)(640) - (-288)^2}}$$

$$= \frac{15600 + 55296}{\sqrt{129600 - 36864}\sqrt{192000 - 82944}}$$

$$= \frac{70896}{\sqrt{92736}\sqrt{109056}} = \frac{70896}{304.526 \times 330.236}$$

$$= \frac{70896}{100565.44} = +0.705.$$

Thus, it is a case of high degree of positive correlation between height and weight.

Illustration 29. The following table gives the distribution of production and also the relatively defective items among them, according to size-groups. Is there any correlation between size and defect in quality.

Size-groups	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

(MBA, IGNOU, 2008)

Solution : Let us find the percentage of defective items and then find correlation between size and defect in quality.

Size-groups	*m.p.* m	$(m-17.5)$ d_x	d_x^2	*No. of items*	*No. of def. items*	*% of def.*	$(Y-50)/5$ d_y	d_y^2	$d_x d_y$
15–16	15.5	−2	4	200	150	75	+5	25	−10
16–17	16.5	−1	1	270	162	60	+2	4	−2
17–18	17.5	0	0	340	170	50	0	0	0
18–19	18.5	+1	1	360	180	50	0	0	0
19–20	19.5	+2	4	400	180	45	−1	1	−2
20–21	20.5	+3	9	300	120	40	−2	4	−6
		$\Sigma d_x = 3$	$\Sigma d_x^2 = 19$			$\Sigma d_y = 4$	$\Sigma d_y^2 = 34$		$\Sigma d_x d_y = -20$

$$r = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

Substituting the values :

$$r = \frac{6(-20) - (3)(4)}{\sqrt{6(19) - (3)^2}\sqrt{6(34) - (4)^2}}$$

$$= \frac{-120 - 12}{\sqrt{105}\sqrt{188}} = \frac{-132}{10.25 \times 13.71} = -0.94$$

There is a very high degree of negative correlation between size and defect in quality.

Illustration 30. Calculate the rank correlation coefficient for the following data giving ranks awarded by two judges to 10 participants in a musical contest :

Rank by Judge I	:	3	5	4	8	9	7	1	2	6	10
Rank by Judge II	:	4	6	3	9	10	7	2	1	5	8

(MBA, Madurai -Kamaraj Univ., 2003)

Solution : CALCULATION OF RANK CORRELATION COEFFICIENT

Participants	*Rank by Judge I* (R_1)	*Rank by Judge II* (R_2)	$(R_1 - R_2)^2$ D^2
A	3	4	1
B	5	6	1
C	4	3	1
D	8	9	1
E	9	10	1
F	7	7	0

G	1	2	1
H	2	1	1
I	6	5	1
J	10	8	4
			$\Sigma D^2 = 12$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 12}{10^3 - 10} = 1 - \frac{72}{990} = 1 - 0.073 = 0.927$$

Illustration 31. Find the rank correlation from the following data :

Candidate	1	2	3	4	5	6	7
Marks awarded by Judge I	86	59	64	74	48	70	94
Marks awarded by Judge II	90	45	72	64	59	60	80

(MBA, Bharthidasan Univ., 2003)

Solution : Since ranks are not given, we first assign ranks and then calculate the rank correlation coefficient.

Candidate	*Marks by Judge I*	R_1	*Marks by Judge II*	R_2	$(R_1-R_2)^2$ D^2
1	86	6	90	7	1
2	59	2	45	1	1
3	64	3	72	5	4
4	74	5	64	4	1
5	48	1	59	2	1
6	70	4	60	3	1
7	94	7	80	6	1
$N=7$					$\Sigma D^2 = 10$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 10}{7^3 - 7} = 1 - \frac{60}{336} = 1 - 0.179 = 0.821$$

Illustration 32. Calculate the correlation coefficient between price and sales from the following data :

Price (Rs.)	100	90	85	92	90	84	88	90
Sales ('00)	5	6	7	6	7	8	8	7

(MBA, Madras Univ., 2003)

Solution : CALCULATION OF CORRELATION COEFFICIENT

Price (Rs.) X	$(X-90)$ d_x	d_x^2	*Sales* Y	$(Y-7)$ d_y	d_y^2	$d_x d_y$
100	+10	100	5	−2	4	−20
90	0	0	6	−1	1	0
85	−5	25	7	0	0	0
92	+2	4	6	−1	1	−2
90	0	0	7	0	0	0
84	−6	36	8	+1	1	−6
88	−2	4	8	+1	1	−2
90	0	0	7	0	0	0
$\Sigma X = 719$	$\Sigma d_x = -1$	$\Sigma d_x^2 = 169$	$\Sigma Y = 54$	$\Sigma d_y = -2$	$\Sigma d_y^2 = 8$	$\Sigma d_x d_y = -30$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$N = 8$, $\Sigma d_x d_y = -30$, $\Sigma d_x = -1$, $\Sigma d_y = -2$, $\Sigma d_x^2 = 169$, $\Sigma d_y^2 = 8$

$$r = \frac{8(-30) - (-1)(-2)}{\sqrt{8(169) - (-1)^2}\sqrt{8(8) - (-2)^2}} = \frac{-240 - 2}{\sqrt{1352 - 1}\sqrt{64 - 4}}$$

$$= \frac{-242}{\sqrt{1351 \times 60}} = \frac{-242}{284.71} = -0.85$$

There is a high degree of negative correlation between price and sales.

Illustration 33. Newspapers in India are complaining that rising level of unemployment is affecting the level of crime in the country. To study this claim, a research team studied a random sample of 12 states in the country. For each state, they measured the level of unemployment rate and the crime rate in the state. Then they did a ranking X = level of unemployment, Y = crime rate, the results are shown in the following table. Higher X ranks more unemployment, and higher Y ranks means higher crime rate. Test the claim of Newspapers.

States	:	1	2	3	4	5	6	7	8	9	10	11	12
Level of unemployment (X)	:	5	8	3	2	6	1	10	12	7	4	9	11
Crime Rate (Y)	:	8	6	9	12	7	10	2	1	5	11	4	3

(MBA, Delhi Univ., 2009)

Solution : For testing the claim of the newspapers, we calculate the rank correlation coefficient.

CALCULATION OF RANK CORRELATION COEFFICIENT

States	R_x	R_y	$(R_x - R_y)^2$ D^2
1	5	8	9
2	8	6	4
3	3	9	36
4	2	12	100
5	6	7	1
6	1	10	81
7	10	2	64
8	12	1	121
9	7	5	4
10	4	11	49
11	9	4	25
12	11	3	64
$N = 12$			$\Sigma D^2 = 558$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$\Sigma D^2 = 558, \quad N = 12$$

$$R = 1 - \frac{6 \times 558}{12^3 - 12} = 1 - \frac{3348}{1728 - 12} = 1 - 1.951 = -0.951$$

There is a high degree of negative correlation between level of unemployment and crime rate.

Illustration 34. Compute Spearman's rank correlation for the following observations :

Candidate	:	1	2	3	4	5	6	7	8
Judge X	:	20	22	28	23	30	30	23	24
Judge Y	:	28	24	24	25	26	27	32	30

(MBA, GGSIP Univ., 2009)

Solution : CALCULATION OF SPEARMAN'S RANK CORRELATION

Candidate	*Judge X*	R_1	*Judge Y*	R_2	$(R_1 - R_2)^2$ D^2
1	20	1	28	6	25.00
2	22	2	24	1.5	0.25
3	28	6	24	1.5	20.25
4	23	3.5	25	3	0.25
5	30	7.5	26	4	12.25
6	30	7.5	27	5	6.25
7	23	3.5	32	8	20.25
8	24	5	30	7	4.00
$N = 8$					$\Sigma D^2 = 88.50$

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\right]}{N^3 - N}$$

$$R = 1 - \frac{6\left[88.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]}{N^3 - N}$$

$$= 1 - \frac{6[88.5 + 0.5 + 0.5 + 0.5]}{504} = 1 - \frac{540}{504} = 1 - 1.071 = -0.071$$

PROBLEMS

1-A : Answer the following questions, each question carries **one** mark:

(*i*) What are the properties of correlation coefficient ?

(*ii*) What are the limitations of correlation analysis ? *(MBA, Madurai-Kamaraj Univ., 2005)*

(*iii*) State the formula for coefficient of correlation in terms of regression coefficients.

(*iv*) What is meant by correlation ?

(*v*) What are the limits of coefficient of correlation ?

(*vi*) What is the use of scatter diagram ?

(*vii*) What is 'Rank correlation' ? *(MBA, Madurai-Kamaraj Univ., 2003)*

(*viii*) Write down the formula for rank correlation coefficient.

(*ix*) Interpret the following value of r : $r = 0$, $r = -1$, $r = +1$, $r = 0.25$.

(*x*) How can 'r' be determined through regression coefficients ?

1-B : Answer the following questions, Each question carries **four** marks:

(*i*) The coefficient of correlation between the variables x and y is 0.64, their covariance is 16. The variance of x is 9. Find the standard deviation of y.

(*ii*) Briefly explain the various types of correlation. *(M.Com., M.K. Univ., 2002)*

(*iii*) What do you understand by correlation ? Describe the uses of the study of correlation. *(M.A. Eco., M.K. Univ., 2003)*

(*iv*) Define correlation between two variables. How is the value of 'r' interpreted ? *(MBA, Madras Univ., 2003)*

(*v*) Does correlation always signify a cause and effect relationship between variables ?

2. (*a*) Explain the meaning and significance of the term correlation.

(*b*) What is correlation ? Clearly explain with suitable illustration its role in taking some business problem. *(MBA, Delhi Univ., 2002)*

3. Define the coefficient of correlation. What is it intended to measure ? How would you interpret the sign and magnitude of a calculated r ? Consider in particular the values of $r = 0$, $r = +1$ and $r = -1$.

4. What is a scatter diagram? How does it help in studying the correlation between two variables, in respect of both its direction and degree ? *(MBA, Delhi Univ., 2007)*

5. (*a*) What is Spearman's rank correlation coefficient? Bring out its usefulness. How does the coefficient differ from Karl Pearson's coefficient of correlation ?

(*b*) Explain briefly the different methods of measuring correlation.

6. (*a*) Does correlation always signify a cause and effect relationship between the variables?

(*b*) Does a high positive correlation between the increase in cigarette smoking and the increase in lung cancer prove that one causes the other ?

7. (*a*) Define correlation coefficient 'r' and give its limits. What interpretation would you give if told that the correlation between the number of truck accidents per year and the age of the driver is (–) 0.60 if only drivers with at least one accident are considered ?

(*b*) What is a scatter diagram ? How do you interpret a scatter diagram ?

8. (*a*) What is correlation? Does it always signify cause and effect relationship?

(*b*) What is coefficient of Rank correlation? Bring out its usefulness. How does this coefficient differ from coefficient of correlation ?

9. (*a*) Prove that the correlation coefficient is unaffected by the change of origin and scale.

(*b*) How is Scatter Diagram helpful in the study of correlation ?

10. (*a*) Explain how covariance of X and Y is related to the coefficient of simple correlation between X and Y.

(*b*) What is meant by correlation ? Distinguish between positive, negative and zero correlation. *(MBA, Delhi Univ., 2005; MBA, UP Tech. Univ., 2006)*

(*c*) Explain critically any two methods of measuring correlation.

11. Find Karl Pearson's coefficient of correlation from the following index numbers and interpret it :

Wages	:	100	101	103	102	104	99	97	98	96	96
Cost of living	:	98	99	99	97	95	02	95	94	90	91

[$r = 0.85$]

12. Find Karl Pearson's coefficient of correlation between capital employed and profit obtained from the following data :

Capital employed (*Rs. crore*)	*Profits obtained* (*Rs. crore*)	*Capital employed* (*Rs. crore*)	*Profits obtained* (*Rs. crore*)
10	2	60	15
20	4	70	14
30	8	80	20
40	5	90	22
50	10	100	50

[$r = 0.85$]

13. Using the following data :

(*a*) Calculate the coefficient of correlation.

(*b*) Estimate the percentage of the group with lung cancer in a country where 15 per cent of the group smoke heavily :

Country	*% of group smoking heavily*	*% of group with lung cancer*
A	10	5
B	20	15
C	20	20
D	30	25
E	30	20

[$r = 0.91$]

14. From the following data, calculate coefficient of correlation between the percentage yield on securities and wholesale price indices for certain years :

Year	:	2004	2005	2006	2007	2008	2009	2010
% Yield on securities	:	5.0	5.1	5.2	4.9	4.8	5.3	5.4
Index No. of wholesale prices	:	140	138	126	132	140	135	132

What inference do you draw from the result ?

[$r = -0.16$]

15. Find the correlation by Karl Pearson's method between the two kinds of assessment of postgraduate student's performance (marks out of 100) :

Roll No. of students	:	1	2	3	4	5	6	7	8	9	10
Internal assessment	:	45	62	67	32	12	38	47	67	42	85
External assessment	:	39	48	65	32	20	35	45	77	30	62

[$r = 0.88$]

16. Two housewives, Mrs. Neena and Mrs. Meena, asked to express their preferences for different kinds of detergents, gave the following replies :

Detergent	:	A	B	C	D	E	F	G	H	I	J
Neena	:	1	2	4	3	7	8	6	5	9	10
Meena	:	1	4	2	3	5	7	6	8	9	10

To what extent the preferences of these two ladies go together ?

[$R = +0.89$]

17. An office contains 10 clerks. The longer-serving clerks feel that they should have a seniority increment based on length of service built into their salary structure. An assessment of their efficiency by their departmental manager and the personnel department produces a ranking of efficiency. This is shown below together with a ranking of their length of service. Do the data support the clerk's claim for seniority increment ?

Ranking according to length of service	:	1	2	3	4	5	6	7	8	9	10
Ranking according to efficiency	:	2	5	3	10	6	4	8	9	7	1

[R = +0.164]

18. The following table gives the frequency, according to age groups, of marks obtained by 68 students in a general knowledge test. Measure the degree of relationship between age and general knowledge.

	Age in Years			
Test Marks	21	22	23	24
200 – 250	4	4	2	1
250 – 300	3	5	4	2
300 – 350	2	6	8	5
350 – 400	1	5	6	10

[$r = 0.415$]

19. Find coefficient of correlation between output and cost per scooter from the following data :

Output of scooter (in '000s) :	3.5	4.0	5.2	6.3	6.8	7.4	8.5	9.0
Cost per scooter (in '000 Rs.) :	12.0	11.8	11.2	10.6	10.3	9.8	9.3	9.2

[r=0.0996]

20. Find the coefficient of correlation between price and sales from the following data :

Price (Rs.) :	103	98	85	92	90	84	88	90	93	95
Sales (Units) :	500	610	700	630	670	800	800	750	700	680

[$r = 0.85$]

21. Calculate correlation coefficient from the following two-way table, with X representing the average salary of families selected at random in a given area and Y representing the average expenditure on entertainment (movies, magazines, etc.) :

Expenditure on entertainment (in 00's Rs.)	*Average salary (in 00's Rs.)*				
	100 – 150	150 – 200	200 – 250	250 – 300	300 – 350
0 – 10	5	4	5	2	4
10 – 20	2	7	3	7	1
20 – 30	–	6	–	4	5
30 – 40	8	–	4	–	8
40 – 50	–	7	3	5	10

[$r = 0.205$]

(MBA, Delhi Univ., 2003)

22. A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so that the students in a pair have approximately equal scores on an intelligence test. In each pair, one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below :

Pair :	1	2	3	4	5	6	7	8	9	10	11
A :	24	29	19	14	30	19	27	30	20	28	11
B :	37	35	16	26	23	27	19	20	16	11	21

(*i*) Find the correlation coefficient between the two sets of scores.

(*ii*) Find the rank correlation coefficient.

(MBA, HPU, 2004)

[(*ii*) –0.175]

23. The mileage (Y) that can be obtained from a certain gasoline depends on the amount (X) of certain chemical in the gasoline. The value of ten observations, where X and Y are measured in appropriate units are shown in the table below :

Amount (X)	*Mileage* (Y)	*Amount* (X)	*Mileage* (Y)
0.10	10.98	0.60	14.63
0.20	11.14	0.70	15.66
0.30	13.17	0.80	13.71
0.40	13.34	0.90	15.43
0.50	14.39	1.00	18.36

Find the coefficient of correlation between X and Y and represent the data by a graph.

24. Calculate the coefficient of correlation between age and sum assured from the data given below and comment on the value :

	Sum assured (in lakh Rs.)				
Age	5	10	15	20	Total
20 – 30	2	3	4	6	15
30 – 40	–	2	3	5	10
40 – 50	–	2	2	3	7
50 – 60	5	8	3	2	18
Total	7	15	12	16	50

[$r = 0.3442$]

(MBA, Delhi Univ., 2002)

25. Compute the coefficient of correlation between dividends and prices of securities as given below :

Security Prices (in Rs.)	*Annual Dividends (in hundred Rs.)*					
	6 – 8	8 – 10	10 – 12	12 – 14	14 – 16	16 – 18
130 – 140	—	—	1	3	4	2
120 – 130	—	1	3	3	3	1
110 – 120	—	1	2	3	2	—
100 – 110	—	2	3	2	—	—
90 – 100	2	2	1	1	—	—
80 – 90	3	1	1	—	—	—
70 – 80	2	1	—	—	—	—

[$r = +0.71$]

26. The top executives of Sonal Electrical rank managerial candidates on the basis of what they know about each candidate. In order to determine if there is any consistency in the ranking obtained in this manner, two vice-presidents were asked to rank the same ten candidates. Compute the coefficient of rank correlation from the following two sets of ranks :

Candidate	A	B	C	D	E	F	G	H	I	J
V-P 1 :	3	1	8	1	4	9	5	7	10	6
V-P 2 :	2	5	9	1	6	10	3	4	8	7

($R = +0.746$]

27. Seven methods of imparting business education were ranked by the MBA students of two universities as follows :

Method of teaching	:	I	II	III	IV	V	VI	VII
Rank by Students of Univ. A	:	2	1	5	3	4	7	6
Rank by Students of Univ. B	:	1	3	2	4	7	5	6

Calculate rank correlation coefficient and comment on its value.

[$R = +0.5$] *(MBA, South Gujarat Univ., 2002; MBA, Delhi Univ, 2005)*

28. (*a*) Coefficient of correlation between X and Y for 20 items is 0.3, mean of X is 15 and that of Y=20, standard deviations are 4 and 5 respectively. At the time of calculation one item 26 was wrongly taken as 17 in case of X series and 35 instead of 30 in case of Y series. Find the correct value of correlation coefficient.

[Correct value of correlation coefficient is 0.504.]

(*b*) In order to find the correlation coefficient between two variables X and Y from 12 pairs of observations, the following calculations were made :

$\Sigma X = 30, \Sigma Y = 5, \Sigma X^2 = 670, \Sigma Y^2 = 285, \Sigma XY = 334.$

On subsequent verification it was found that the pair ($X = 11, Y = 4$) was copied wrongly, the correct value being ($X = 10, Y = 14$). Find the correct value of correlation coefficient.

[$r = 0.78$]

29. A Statistician while calculating the correlation coefficient between two variates 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	:	6	8
Y	:	14	6

While the correct values were

X	:	8	16
Y	:	12	8

Obtain the correct value of the correlation coefficient.

(M.Com., Madras Univ., 2009)

[$r = 0.67$]

30. The following data relate to the prices and supplies of a commodity during a period of eight years :

Price (Rs./kg) :	10	12	18	16	15	19	18	17
Supply (100 kg) :	30	35	45	44	42	48	47	46

Calculate the coefficient of correlation between the two series.

(MBA, Punjab Univ., 2002)

[$r = 0.98$]

31. Calculate the coefficient of correlation between family income and its percentage spent on food for the following data :

Family Income	*Food Expenditure (in percentage)*				
(in Rs.)	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
12000 – 13000	2	3	1	4	—
13000 – 14000	3	4	2	1	5
14000 – 15000	4	1	5	12	8
15000 – 16000	1	2	3	—	4
16000 – 17000	5	6	—	3	1

[$r = 0.1048$]

32. Calculate the coefficient of correlation and probable error of r between the values of X and Y given below :

X	:	78	98	96	69	59	79	68	61
Y	:	125	137	156	112	107	136	123	108

[$r = 0.955$, P.E.r. = 0.021]

(M.Com., Sukhadia Univ., 2000)

33. Find the coefficient of correlation for the following bivariate frequency distribution :

Marks in Physics	*Marks in Mathematics*						
	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99	Total
90 – 99				2	4	4	10
80 – 89			1	4	6	5	16
70 – 79			5	10	8	1	24
60 – 69	1	4	9	5	2		21
50 – 59	3	6	6	2			17
40 – 49	3	5	4				12
Total	7	15	25	23	20	10	100

[$r = +0.769$]

(M. Com., M.D. Univ., 2003)

34. The bivariate frequency distribution based on monthly salary and age of 100 employees working in some large-scale commercial organisation is as under :

	Monthly Salary (in 000's Rs.)			
Age (Years)	8 – 10	10 – 12	12 – 14	14 – 16
20 and less than 30	16	6	—	—
30 and less than 40	4	10	4	4
40 and less than 50	—	4	18	12
50 and less than 60	—	—	10	12

Compute Karl Pearson's coefficient of correlation between age and monthly salary of employees and comment on its value.

[$r = +0.763$]

35. A survey regarding income and savings provided the following data :

	Saving (Rs.)			
Income (Rs.)	1000	2000	3000	4000
8000	8	4	—	—
12000	—	12	24	6
16000	—	9	7	2
20000	—	—	10	5
24000	—	—	9	4

Compute Karl Pearson's coefficient of correlation and interpret its value.

[$r = +0.522$]

(MBA Delhi Univ., 2006)

36. Calculate the coefficient of correlation from the following data and interpret the value.

Advertising expenditure (Rs. lakhs)	:	10	12	13	23	27	30
Sales turnover (Rs. crores)	:	40	42	46	48	50	56

[$r = +0.956$]

(MBA, Delhi Univ., 2002)

37. You are given the following data of marks obtained by 11 students in statistics in two tests, one before and the other after special coaching :

First Test (Before coaching)	:	23	20	19	21	18	20	18	17	23	16	19
Second Test (After coaching)	:	24	19	22	18	20	22	20	20	23	20	17

Do the marks indicate that the special coaching has benefited the students ?

[$r = +0.477$]

(M.Com., Delhi Univ., 2000)

38. The scores of students in an examination in Mathematics and Statistics are given below :

Student No.	:	1	2	3	4	5	6	7	8
Marks in Mathematics	:	70	48	58	55	54	50	60	52
Marks in Statistics	:	62	47	53	60	55	68	51	48

Find : (*i*) Correlation coefficient, and

(*ii*) Rank correlation coefficient and compare the two values.

[(*i*) $r = 0.246$, (*ii*) $r = 0.286$]

39. The following data show the marks of 10 students in Mathematics and Statistics in an examination :

Marks in Mathematics :	45	70	65	30	90	40	50	75	85	60
Marks in Statistics :	35	90	70	40	95	40	60	80	80	50

Find Karl Pearson's coefficient of correlation and its probable error. *(MBA, Vikram Univ., 2007)*

40. A researcher collected the following information for two variables x and y :

No. of Pairs = 20, $r = 0.5$, $\bar{x} = 15$, $\bar{y} = 20$, $\sigma_x = 4$, $\sigma_y = 5$

Later it was found that one pair of value has been wrongly taken as $\frac{x \mid y}{16 \mid 30}$ whereas the correct values were $\frac{x \mid y}{26 \mid 35}$. Find the correct value of r.

[$r = 0.559$] *(MBA, MD Univ., 2002)*

41. Calculate the Karl Pearson's Coefficient of Correlation between age and playing habits from the data given below. Comment on the value :

Age	:	20	21	22	23	24	25
No. of students	:	500	400	300	240	200	160
Regular players	:	400	300	180	96	60	24

[$r = -0.991$] *(MBA, Delhi Univ., 2006)*

42. The following bivariate frequency distribution relates to the age and salary of 100 computer operators working in an organisation. Find the coefficient of correlation and interpret its value.

	Salary (Rs.)			
Age (Yrs)	*15000 – 16000*	*16000 – 17000*	*17000 – 18000*	*18000 – 19000*
20 – 30	4	6	5	2
30 – 40	2	5	8	5
40 – 50	8	12	20	2
50 – 60	—	8	12	1

[$r = 0.057$] *(MBA, Delhi Univ., 2006)*

43. Compute the rank correlation coefficient from the following data :

Series X :	115	109	112	87	98	98	120	100	98	118
Series Y :	75	73	85	70	76	65	82	73	68	80

[$R = 0.33$] *(MBA, KU, 2006)*

44. From the following data calculate coefficient of correlation between age and playing habit. How do you interpret the result?

Age group	*No. of Employees*	*No. of regular players*
20 – 30	50	20
30 – 40	120	60
40 – 50	80	24
50 – 60	40	4
60 – 70	20	1

(MBA, Guru Jambheshwar Univ., 2005)

45. Calculate the coefficient of correlation from the following data :

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

[$r = +0.603$] *(MBA, Madurai-Kamaraj Univ., 2006)*

46. Calculate the coefficient of correlation from the following data :

X :	100	200	300	400	500	600	700
Y :	30	50	60	80	100	110	130

[$r = 0.997$]

47. Find the coefficient of correlation for the following :

A :	5	10	5	11	12	4	3	2	7	1
B :	1	6	2	8	5	1	4	6	5	2

(MBA, Madurai-Kamaraj Univ., 2003)

48. Two designs A and B gave the following output in 9 trails of each. Which is a better design ? Why ?

A :	16	16	53	15	31	17	14	30	20
B :	18	27	23	21	22	26	39	17	28

49. Calculate Pearson's coefficient of correlation from the following taking 100 and 50 as the assumed average of X and Y respectively :

X :	104	111	104	114	118	117	105	108	106	100	104	105
Y :	57	55	47	45	45	50	64	63	66	62	69	61

(MBA, Bharathidasan Univ., 2001)

50. Find the correlation coefficient from the following data :

X :	53	25	19	37	42	10	15
Y :	9	6	5	7	7	4	5

[$r = 0.348$]

51. The marking of 10 trainees in two skills, programming and analysis are as follows. What is the coefficient of rank correlation?

Programming :	3	5	8	4	7	10	2	1	6	9
Analysis :	6	4	9	8	1	2	3	10	5	7

[$r = -0.297$]

(MBA, Bharathidasan Univ., 2006)

52. The GE Capital is in the business of making bids on investments offered by various firms that desire additional financing. The company has collected the following data on yearly investments and interest rates :

Year :	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*
Yearly Investments (Thousand of Rs.) :	1080	948	920	1119	1695	2150	2170	2230	1880	1425
Average Interest Rate (%) :	4.8	5.1	5.9	5.1	4.8	3.8	3.7	4.5	4.9	6.2

Is the relationship between these variables significant? If the average interest rate is 6% five years from now, can yearly investment be forecast?

(MBA, Delhi Univ., 2009)

53. A consulting firm is preparing a study on consumer behaviour. The company collected the following data in thousand dollars to determine whether there is a relationship between consumer income and consumption levels :

Consumer No. :	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Income :	24.3	12.5	31.2	28.0	35.1	10.5	23.2	10.0	8.5	15.9	14.7	15
Consumption :	16.2	8.5	15	17	24.2	11.2	15	7.1	3.5	11.5	10.7	9.2

(*a*) Calculate correlation coefficient for the above data.

(*b*) Compute and interpect the regression model. Tell about the relationship between consumption and income? What consumption would the model predict for someone who earns $27500?

(MBA, Delhi Univ., 2009)

*****

8

Regression Analysis

INTRODUCTION

After having established the fact that two variables are closely related, we may be interested in estimating (predicting) the value of one variable given the value of another. For example, if we know that advertising and sales are correlated, we may find out the expected amount of sales for a given advertising expenditure or the required amount of expenditure for achieving a fixed sales target. *The statistical tool with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable is called* **regression.** With the help of regression analysis, we are in a position to find out the *average probable* change in one variable given a certain amount of change in another.

The dictionary meaning of the term 'regression' is the act of returning or going back. The term 'regression' was first used in 1877 by Francis Galton while studying the relationship between the heights of fathers and sons. His study of height of about one thousand fathers and sons revealed a very interesting relationship, *i.e.*, tall fathers tend to have tall sons and short fathers, short sons; but the average height of the sons of a group of tall fathers is less than that of the tall fathers and the average height of the sons of a group of short fathers is greater than that of the short fathers. The line describing this tendency to regress or going back was called by Galton a 'Regression Line'. The term is still used to describe the line drawn for a group of points to represent the trend present, but it no longer necessarily carries the original implication that Galton intended. These days there is a growing tendency of the modern writers to use the term *estimating line or predicting line* instead of *regression line.*

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. For example, if we know that two variables price (X) and demand (Y) are closely related we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of a commodity are closely related, we can find out the expected price for a certain amount of tax levy. The regression analysis helps in three important ways :

1. It provides estimates of values of the dependent variables from values of independent variables. The device used to accomplish the estimation procedure is the regression line which describes the average relationship existing between X and Y variables.

2. The second goal of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimations. For this purpose, the standard error of estimate is calculated. If the line fits the data closely, that is, if there is relatively little scatter of the observations around the regression line, good estimate can be made of Y variable. On the other hand, if there is a great deal of scatter of the observations around the fitted regression line, the line will not produce accurate estimates of the dependent variable.

3. With the help of regression analysis, we can obtain a measure of the degree of association or correlation that exists between the two variables. The coefficient of determination calculated for this purpose measures the strength of the relationship that exists between the variables. It assesses the proportion of variance that has been accounted for by the regression equation.

The tool of regression analysis can be extended to three or more variables. But in this text we shall confine ourselves to the problems of two variables only, *i.e.*, simple regression.

Difference between Correlation and Regression Analysis

There are two important points of difference between correlation and regression analysis :

1. Whereas correlation coefficient is a measure of degree of relationship between X and Y, the objective of regression analysis is to study the '*nature of relationship*' between the variables.

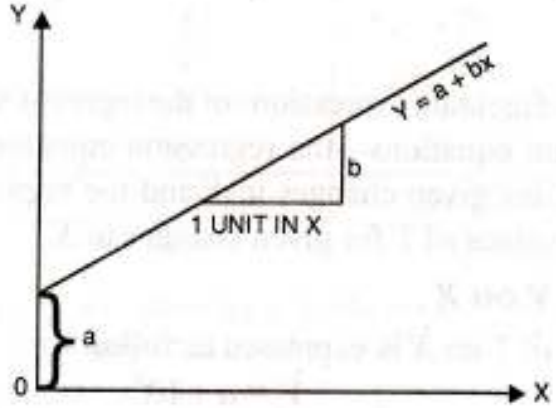
2. The cause and effect relation is clearly indicated through regression analysis than by correlation. Correlation is merely a tool of ascertaining the degree of relationship between two variables and, therefore, we cannot say that one variable is the cause and the other the effect.

THE LINEAR BIVARIATE REGRESSION MODEL

In regression analysis, as in other types of statistical studies, we usually proceed by observing the sample data and using the results obtained as estimates of the corresponding population relationship. To make valid inferences, we must assume some population model. For a bivariate population, there are many possible models that can be constructed to describe the mutual variations of the two variables. The particular one in which we are interested is called the *simple linear regression model*. This model is constructed under the following set of assumptions :

(1) The value of the dependent variable, Y, is dependent in some degree upon the value of the independent variable, X. The dependent variable is assumed to be a random variable, but the values of X are assumed to be fixed quantities that are selected and controlled by the experimenter. The requirement that the independent variable assumes fixed values, however, is not a critical one. Useful results can still be obtained by regression analysis in the case where both X and Y are random variables.

(2) The average relationship between X and Y can be adequately described by a linear equation $Y = a + bX$ whose geometrical presentation is a straight line as in the diagram that follows :



As is clear from the above diagram, the height of the line tells the average value of Y at a fixed value of X. When $X = 0$, the average value of Y is equal to a. The value of a is called the Y intercept, since it is the point at which the straight line crosses the Y-axis. The slope of the line is measured by b, which gives the average amount of change of Y per unit change in the value of X. The sign of b also indicates the type of relationship between Y and X.

(3) Associated with each value of X there is a sub-population of Y. The distribution of the sub-population may be assumed to be normal or non-specified in the sense that it is unknown. In any event, the distribution of each population Y is conditional to the value of X.

(4) The mean of each sub-population Y is called the expected value of Y for a given X : $E(Y/X) = \mu_{yx}$. Furthermore, under the assumption of a linear relationship between X and Y, all values of $E(Y/X)$ or μ_{yx} must fall on a straight line. That is

$$E(Y/X) = \mu_{yx} = a + bX$$

which is the population regression equation for our bivariate linear model. In this equation a and b are called the population regression coefficients.

(5) An individual value in each sub-population Y, may be expressed as :

$$Y = E(Y/X) + e$$

where e is the deviation of a particular value of Y from μ_{yx} and is called the *error term* or *the stochastic disturbance term*. The errors are assumed to be independent random variables because Y's are random variables and independent. The expectations of these errors are zero; $E(e) = 0$. Moreover, if Y's are normal variables, the error can also be assumed to be normal.

(6) It is assumed that the variances of all sub-populations, called variances of the regression, are identical.

Regression Lines

If we take the case of two variables X and Y, we shall have two regression lines as the regression line of X on Y and the regression line of Y on X. The regression line of Y on X gives the most probable values of Y for given values of X and the regression line of X on Y gives the most probable values of X for given values of Y. Thus, we have two regression lines. However, when there is either perfect positive or perfect negative correlation between the two variables, the two regression lines will coincide, *i.e.*, we will have one line. The farther the two regression lines are from each other, the lesser is the degree of correlation and the nearer the two regression lines to each other, the higher is the degree of correlation. If the variables are independent, r is zero and the lines of regression are at right angles, *i.e.*, parallel to X-axis and Y-axis.

It should be noted that the regression lines cut each other at the point of average of X and Y, *i.e.*, if from the point where both the regression lines cut each other, a perpendicular is drawn on the X-axis, we will get the mean value of X and if from the point a horizontal line is drawn on the Y-axis, we will get the mean value of Y.

Regression Equations

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations—the regression equation of X on Y is used to describe the variations in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Regression Equation of Y on X

The regression equation of Y on X is expressed as follows :

$$Y_e = a + bX$$

Where Y_e is the dependent variable to be estimated and X is the independent variable.

In this equation a and b are two unknown constants (fixed numerical values) which determine the position of the line completely. The constants are called the parameters of the line. If the value of either or both of them is changed, another line is determined. The parameter 'a' determines the *level* of the fitted line (*i.e.*, the distance of the line directly above or below the origin). The parameter 'b' determines the *slope* of the line, *i.e.*, the change in Y for unit change in X.

If the values of the constants 'a' and 'b' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the *method of least squares*

which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the vertical deviations of the actual Y values from the estimated Y values is the least, or, in other words, in order to obtain a line which fits the points best, $(Y - Y_c)^2$ should be minimum*. Such a line is known as the line of best fit.

With a little algebra and differential calculus, it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters a and b such that the least squares requirement is fulfilled.

$$\Sigma Y = Na + b\Sigma X \quad \ldots (i)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \ldots (ii)$$

These equations are usually called the *normal equations*. In the equations ΣX, ΣY, ΣXY, ΣX^2 indicate totals which are computed from the observed pairs of values of two variables X and Y to which the least squares estimating line is to be fitted and N is the total number of observed pairs of values.

Regression Equation of X on Y

The regression equation of X on Y is expressed as follows :

$$X = a + bY$$

To determine the values of a and b the following two normal equations are to be solved simultaneously :

$$\Sigma X = Na + b\Sigma Y \quad \ldots (i)$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2 \quad \ldots (ii)$$

Illustration 1. Calculate the regression equations of X on Y and Y on X from the following data :

X	:	1	2	3	4	5
Y	:	2	5	3	8	7

Solution :

CALCULATION OF REGRESSION EQUATIONS

X	Y	X^2	Y^2	XY
1	2	1	4	2
2	5	4	25	10
3	3	9	9	9
4	8	16	64	32
5	7	25	49	35
$\Sigma X = 15$	$\Sigma Y = 25$	$\Sigma X^2 = 55$	$\Sigma Y^2 = 151$	$\Sigma XY = 88$

*$\Sigma (Y - Y_c)^2$ should be minimum or $\Sigma (Y - a - bX)^2$ should be minimum (since $Y_c = a + bX$).

Let $S = \Sigma (Y - a - bX)^2$

Differentiating partially with respect to a and b

$$\frac{\partial S}{\partial a} = \Sigma (Y - a - bX)(-1) = 0$$

$$\frac{\partial S}{\partial b} = \Sigma (Y - a - bX)(-X) = 0$$

or $$\Sigma (Y - a - bX) = 0$$

$$\Sigma (Y - a - bX)X = 0$$

or $$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Regression equation of X on Y is given by

$$X = a + bY$$

The normal equations are :

$$\Sigma X = Na + b\Sigma Y$$
$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values, we get

$$15 = 5a + 25b \quad \ldots(i)$$
$$88 = 25a + 151b \quad \ldots(ii)$$

Solving (i) and (ii), we get

$$a = 0.5 \text{ and } b = 0.5$$

Hence the required regression equation of X on Y is given by

$$X = 0.5 + 0.5Y$$

Regression equations of Y on X is : $Y = a + bX$

The normal equations are:

$$\Sigma Y = Na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values, we get

$$25 = 5a + 15b \quad \ldots(iii)$$
$$88 = 15a + 55b \quad \ldots(iv)$$

Solving (iii) and (iv), we get

$$a = 1.10 \text{ and } b = 1.3$$

Hence the required regression equation of Y on X is given by

$$Y = 1.10 + 1.30X$$

Illustration 2. After investigation it has been found the demand for automobiles in a city depends mainly, if not entirely, upon the number of families residing in that city. Below are given figures for the sales of automobiles in the five cities for the year 2003 and the number of families residing in those cities.

City	*No. of families in lakhs (X)*	*Sale of Automobiles in 000's (Y)*
A	70	25.2
B	75	28.6
C	80	30.2
D	60	22.3
E	90	35.4

Fit a linear regression equation of Y on X by the least square method and estimate the sales for the year 2006 for city A which is estimated to have 100 lakh families assuming that the same relationship holds true.

Solution. CALCULATION OF REGRESSION EQUATION

City	X	Y	X^2	XY
A	70	25.2	4,900	1,764
B	75	28.6	5,625	2,145
C	80	30.2	6,400	2,416
D	60	22.3	3,600	1,338
E	90	35.4	8,100	3,186
	$\Sigma X = 375$	$\Sigma Y = 141.7$	$\Sigma X^2 = 28,625$	$\Sigma XY = 10,849$

Regression equation of Y on X is $Y = a + bX$.

To determine the values of a and b, we shall solve the normal equations

$$\Sigma Y = Na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values from the table, the normal equations become

$$141.7 = 5a + 375b \quad \ldots(i)$$
$$10,849 = 375a + 28,625b \quad \ldots(ii)$$

Multiplying Eqn. (i) by 75 and subtracting from Eqn. (ii), we get

$$221.5 = 500b \text{ or } b = 0.443$$

Substituting the value of b in Eqn. (i), we have

$$24.425 = 5a \text{ or } a = -4.885$$

Therefore, the regression equation of Y on X is

$$Y = -4.885 + 0.443X$$

Estimated sales for the year 2006 for city A

$$Y = -4.885 + 0.443\,(100)$$
$$= -4.885 + 44.3 = 39.415$$

Hence it is expected that about 39,415 autos would be sold in city A having a population of 100 lakh families.

Deviations taken from Arithmetic Means of X and Y

The calculations by the direct method discussed above are quite cumbersome when the values of X and Y are large. The work can be simplified if instead of dealing with the actual values of X and Y we take the deviations of X and Y series from their respective means. In such a case, the equation $Y = a + bX$ is changed to :

$$Y - \bar{Y} = b_{yx}\,(X - \bar{X})$$

The value of b_{yx} can be easily obtained as follows :

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

where $x = (X - \bar{X})$ and $y = (Y - \bar{Y})$

The two normal equations which we had written earlier when changed in terms of x and y become

$$\Sigma y = Na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \ldots (i)$$

Since $\Sigma x = \Sigma y = 0$ [deviations being taken from means] $\ldots (ii)$

Equation (i) reduces to

$$Na = 0, \quad \therefore a = 0$$

Equation (ii) reduces to

$$\Sigma xy = b\Sigma x^2 \quad \therefore b \text{ or } b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

After obtaining the value of b_{yx} the regression equation can easily be written in terms of X and Y by substituting for y, $(Y - \bar{Y})$ and for x, $(X - \bar{X})$.

Similarly, the regression equation $X = a + bY$ is reduced to $(X - \bar{X}) = b_{xy}\,(Y - \bar{Y})$ and the value of b_{xy} can be similarly obtained as

$$b_{xy} = \frac{\Sigma xy}{\Sigma x^2}$$

Illustration 3. In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales :

Salesmen	:	1	2	3	4	5	6	7	8	9	10
Test score	:	40	70	50	60	80	50	90	40	60	60
Sales ('000 Rs.)	:	2.5	6.0	4.0	5.0	4.0	2.5	5.5	3.0	4.5	3.0

Calculate the regression equation of sales on test scores and estimate the probable weekly sales volume if a salesman makes score of 100.

Solution. Let sales be denoted by Y and test scores by X. We have to fit a regression equation of Y on X, *i.e.*, $Y - \bar{Y} = b_{yx}$ $(X - \bar{X})$.

CALCULATION OF REGRESSION EQUATION

Salesmen	Test Score X	$(X-\bar{X})$ x	x^2	Sales Y	$(Y-\bar{Y})$ y	y^2	xy
1	40	–20	400	2.5	–1.5	2.25	+30
2	70	+10	100	6.0	+2.0	4.00	+20
3	50	–10	100	4.0	0	0	0
4	60	0	0	5.0	1.0	1.00	0
5	80	+20	400	4.0	0	0	0
6	50	–10	100	2.5	–1.5	2.25	+15
7	90	+30	900	5.5	+1.5	2.25	+45
8	40	–20	400	3.0	–1.0	1.00	+20
9	60	0	0	4.5	+0.5	0.25	0
10	60	0	0	3.0	–1.0	1.00	0
$N=10$	$\Sigma X=600$	$\Sigma x=0$	$\Sigma x^2=2,400$	$\Sigma Y=40$	$\Sigma y=0$	$\Sigma y^2=14$	$\Sigma xy=130$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{600}{10} = 60; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{10} = 4$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{130}{2,400} = 0.054$$

The regression equation of sales and test scores is given as :

$$Y - 4 = 0.054\,(X - 60)$$

$$Y_c = 0.76 + 0.054\,X$$

When X is 100, Y would be

$$Y = 0.76 + 0.054\,(100) = 6.16.$$

Thus the most probable weekly sales volume if salesman makes a score of 100 is 6.16 thousand rupees.

Deviations taken from Assumed Means

When actual means of X and Y variables are in fractions, the calculations can be simplified by taking the deviations from the assumed mean. The value of b, *i.e.*, the regression coefficient, will be calculated as follows :

Regression equation of X on Y : $(X-\bar{X}) = b_{xy}(Y-\bar{Y})$

where $$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

Regression equation of Y on X : $(Y-\bar{Y}) = b_{yx}(X-\bar{X})$

where $$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

Once the values of b_{xy} and b_{yx} are determined in the above manner, the regression equations can be obtained very easily.

Illustration. 4. A company wants to assess the impact of R & D expenditure on its annual profit. The following table presents the information for the last eight years :

Years	:	2010	2009	2008	2007	2006	2005	2004	2003
R & D expenditure (Rs. '000)	:	9	7	5	10	4	5	3	2
Annual Profit (Rs. '000)	:	45	42	41	60	30	34	25	20

Estimate the regression equation and predict the annual profit for 2009 for an allocated sum of Rs. 100,000 as R & D expenditure.

Solution. Let R & D expenditure be denoted by X and annual profit by Y.

CALCULATION OF REGRESSION EQUATION

Year	X	$(X-6)$ d_x	d_x^2	Y	$(Y-37)$ d_y	d_y^2	$d_x d_y$
2003	2	−4	16	20	−17	289	+68
2004	3	−3	9	25	−12	144	+36
2005	5	−1	1	34	−3	9	+3
2006	4	−2	4	30	−7	49	+14
2007	10	+4	16	60	+23	529	+92
2008	5	−1	1	41	+4	16	−4
2009	7	+1	1	42	+5	25	+5
2010	9	+3	9	45	+8	64	+24
	$\Sigma X = 45$	$\Sigma d_x = -3$	$\Sigma d_x^2 = 57$	$\Sigma Y = 297$	$\Sigma d_y = 1$	$\Sigma d_y^2 = 1125$	$\Sigma d_x d_y = 238$

Fitting regression equation of Y on X, we get

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{297}{8} = 37.125; \qquad \bar{X} = \frac{\Sigma X}{N} = \frac{45}{8} = 5.625$$

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8 \times 238 - (-3)(1)}{8 \times 57 - (-3)^2} = \frac{1904 + 3}{456 - 9} = \frac{1907}{447} = 4.266$$

$Y - 37.125 = 4.266\,(X - 5.625)$ Or $Y - 37.125 = 4.266\,X - 23.996$

$Y = 13.129 + 4.266X$; When X is 10, Y shall be

$Y = 13.129 + 4.266(100) = 439.729$

Thus the likely expenditure on Research and Development for an allocation of Rs. 100,000 is Rs. 439.729.

Regression Coefficients

The Quantity b in the regression equations is called the "regression coefficient" or "slope coefficient". Since there are two regression equations, therefore, there are two regression coefficients—regression coefficient of X on Y and regression coefficient of Y on X.

Regression Coefficient of X on Y

The regression coefficient of X on Y is represented by the symbol b_{xy} or b_1. It measures the amount of change in X corresponding to a unit change in Y. The regression coefficient of X on Y is given by

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

When deviations are taken from the means of X and Y, the regression coefficient is obtained by

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$$

When deviations are taken from assumed means, the value of b_{xy} is obtained as follows :

$$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

Regression Coefficient of Y on X

The regression coefficient of Y on X is represented by b_{yx} or b_2. It measures the amount of change in Y corresponding to a unit change in X. The value of b_{yx} is given by

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

When deviations are taken from actual means of X and Y,

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

When deviations are taken from assumed means,

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

Properties of the Regression coefficients

(1) The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically :

$$r = \sqrt{b_{xy} \times b_{yx}}$$

Proof. $$b_{xy} = r\frac{\sigma_x}{\sigma_y} \; ; b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$\therefore \quad b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2.$$

(2) If one of the regression coefficients is greater than unity, the other must be less than unity, since the value of the coefficient of correlation cannot exceed unity. For example, if $b_{xy} = 1.2$ and $b_{yx} = 1.4$, r would be $\sqrt{1.2 \times 1.4} = 1.29$ which is not possible.

(3) Both the regression coefficients will have the same sign, *i.e.*, they will be either positive or negative. In other words, it is not possible that one of the regression coefficients is having minus sign and the other plus sign.

(4) The coefficient of correlation will have the same sign as that of regression coefficients, *i.e.*, if regression coefficients have a negative sign, r will also have negative sign and if the regression coefficients have a positive sign, r would also be positive. For example,

$$\text{if } b_{xy} = -0.2 \text{ and } b_{yx} = -0.8$$

$$r = -\sqrt{0.2 \times 0.8} = -0.4$$

(5) The average value of the two regression coefficients would be greater than the value of coefficient of correlation. In symbols $(b_{xy} + b_{yx})/2 > r$. For example, if $b_{xy} = 0.8$ and $b_{yx} = 0.4$, the average of the two values would be $(0.8 + 0.4)/2 = 0.6$ and the value of r would be $\sqrt{0.8 \times 0.4} = 0.566$ which is less than 0.6.

(6) Regression coefficients are independent of change of origin but not scale.*

---

*Proof

$$b_{yx} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} \quad \text{or} \quad b_{yx} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}$$

Let $$u = \frac{X - a}{h} \text{ and } v = \frac{Y - b}{k}$$

Then $$X = a + hu, \text{ and } Y = b + kv$$

and $$\bar{X} = a + h\bar{u}\,; \quad \bar{Y} = b + k\bar{v}$$

Subtracting, we get

$$(X - \bar{X}) = h(u - \bar{u})\,; (Y - \bar{Y}) = k(v - \bar{v})$$

Substituting these values in the above formula, we get

$$b_{yx} = \frac{\Sigma hk\,(u - \bar{u})(v - \bar{v})}{\Sigma h^2 (u - \bar{u})^2}$$

$$= \frac{k}{h}\frac{\Sigma(u - \bar{u})(v - \bar{v})}{\Sigma(u - \bar{u})^2} = \frac{k}{h} b_{vu}$$

Similarly, we have $$b_{xy} = \frac{h}{k} b_{uv}$$

Hence the result.

Illustration 5. On the basis of figures recorded below for 'Supply' and 'Price' for nine years, calculate the regression coefficients and the value of r :

Year	:	2002	2003	2004	2005	2006	2007	2008	2009	2010
Supply	:	80	82	86	91	83	85	89	96	93
Price	:	145	140	130	124	133	127	120	110	116

Solution. Let the price be denoted by Y and supply by X.

CALCULATION OF REGRESSION COEFFICIENTS

Year	*Supply* X	$(X-90)$ d_x	d_x^2	*Price* Y	$(Y-127)$ d_y	d_y^2	$d_x d_y$
2002	80	–10	100	145	+18	324	–180
2003	82	– 8	64	140	+13	169	– 104
2004	86	– 4	16	130	+3	9	– 12
2005	91	+1	1	124	–3	9	–3
2006	83	–7	49	133	+ 6	36	–42
2007	85	–5	25	127	0	0	0
2008	89	–1	1	120	–7	49	+ 7
2009	96	+ 6	36	110	– 17	289	–102
2010	93	+ 3	9	116	– 11	121	–33
$N = 9$	$\Sigma X = 785$	$\Sigma d_x = -25$	$\Sigma d_x^2 = 301$	$\Sigma Y = 1{,}145$	$\Sigma d_y = +2$	$\Sigma d_y^2 = 1{,}006$	$\Sigma d_x d_y = -469$

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{9 \times -469 - (-25)(2)}{9 \times 301 - (-25)^2} = \frac{-4221 + 50}{2709 - 625} = -\frac{4171}{2084} = -2.001$$

$$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$= \frac{9 \times -469 - (-25)(2)}{9 \times 1006 - (-2)^2} = \frac{-4221 + 50}{9054 - 4} = -\frac{4171}{9050} = -0.461$$

$$r = \sqrt{b_{yx} \times b_{xy}} = -\sqrt{2.001 \times .461} = -0.96.$$

It is a case of very high degree of negative correlation.

Illustration 6. The following data relate to advertising expenditure (in lakhs of rupees) and their corresponding sales (in crores of rupees) :

Advertising Expenditure	:	10	12	15	23	20
Sales	:	14	17	23	25	21

Estimate (*i*) the sales corresponding to advertising expenditure of Rs. 30 lakhs and (*ii*) the advertising expenditure for a sales target of Rs. 35 crores.

Solution. Let advertising expenditure be denoted by X and sales by Y.

CALCULATION OF REGRESSION EQUATIONS

X	$(X-16)$ x	x^2	Y	$(Y-20)$ y	y^2	xy
10	–6	36	14	–6	36	+36
12	–4	16	17	–3	9	+12
15	–1	1	23	+3	9	–3
23	+7	49	25	+5	25	+35
20	+4	16	21	+1	1	+4
$\Sigma X = 80$	$\Sigma x = 0$	$\Sigma x^2 = 118$	$\Sigma Y = 100$	$\Sigma y = 0$	$\Sigma y^2 = 80$	$\Sigma xy = +84$

(*i*) *Regression equation of Y on X* : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{100}{5} = 20; \quad \bar{X} = \frac{\Sigma X}{N} = \frac{80}{5} = 16$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{84}{118} = 0.712$$

$$Y - 20 = .712\,(X - 16)$$

$$Y - 20 = .712X - 11.392 \text{ or } Y = 8.608 + 0.712X$$

$$Y_{30} = 8.608 + 0.712\,(30) = 8.608 + 21.36 = 29.968$$

Thus the likely sales corresponding to advertising expenditure of Rs. 30 lakhs is Rs. 29.968 crores.

(*ii*) *Regression equation of X on Y* : $X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{84}{80} = 1.05$$

$$X - 16 = 1.05\,(Y - 20)$$

$$X = -5 + 1.05Y$$

$$X_{35} = -5 + 1.05(35) = -5 + 36.75 = 31.75.$$

Thus, the advertising expenditure for a sales target of Rs. 35 crores is Rs. 31.75 lakhs.

Regression Equations in Bivariate Grouped Frequency Distributions

While calculating regression equations for bivariate grouped frequency distributions, first of all we will have to prepare a correlation tables as was discussed in the chapter on Correlation. Then we will find out the value of $\bar{X}$, $\bar{Y}$ and the two regression coefficients and proceed in the usual manner. However, special care must be exercised while calculating the value of regression coefficient because regression coefficients are independent of the change of origin but not of scale. The values of b_{xy} and b_{yx} shall be obtained as follows :

$$b_{xy} = \frac{N\Sigma fd_x d_y - \Sigma fd_x\,\Sigma fd_y}{N\Sigma fd_x^2 - (\Sigma fd_x)^2} \times \frac{h}{k}$$

where h = width of the class-interval of the X variable

and k = width of class-interval of the Y variable.

$$b_{yx} = \frac{N\Sigma fd_x d_y - \Sigma fd_x\,\Sigma fd_y}{N\Sigma fd_x^2 - (\Sigma fd_x)^2} \times \frac{h}{k}$$

(For proof see section on properties of Regression Coefficients.)

Illustration 7. Obtain the two regression equations from the following bivariate frequency distribution :

Sales Revenue (*in Rs. lakhs*)	*Advertising Expenditure* (*in Rs. thousand*)			
	5 – 15	*15 – 25*	*25 – 35*	*35 – 45*
75 – 125	3	4	4	8
125 – 175	8	6	5	7
175 – 225	2	2	3	4
225 – 275	3	3	2	2

Estimate (*i*) the sales corresponding to advertising expenditure of Rs. 50 thousand (*ii*) the advertising expenditure for a sales revenue of Rs. 300 lakhs (*iii*) the coefficient of correlation and interpret its value. (*MBA, Delhi Univ., 2004, 2007*)

Solution. Let sales revenue be denoted by X and advertising expenditure by Y.

CALCULATION OF REGRESSION LINES

X \ Y (m.p. / dy / dx)	m.p.	dx	10 5 – 15 −2	20 15 – 25 −1	30 25 – 35 0	40 35 – 45 +1	f	fd_x	fd_x^2	fd_xd_y
100	75–125	−1	3 | 6	4 | 4	4 | 0	8 | −8	19	−19	19	2
150	125–175	0	8 | 0	6 | 0	5 | 0	7 | 0	26	0	0	0
200	175–225	+1	2 | −4	2 | −2	3 | 0	4 | 4	11	11	11	−2
250	225–275	+2	3 | −12	3 | −6	2 | 0	2 | 4	10	20	40	−14
f			16	15	14	21	$N = 66$	$\Sigma fd_x = 12$	$\Sigma fd_x^2 = 70$	$\Sigma fd_xd_y = -14$
fd_y			−32	−15	0	21	$\Sigma fd_y = -26$			
fd_y^2			64	15	0	21	$\Sigma fd_y^2 = 100$			
fd_xd_y			−10	−4	0	0	$\Sigma fd_xd_y = -14$			

Regression equation of X on Y : $X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$\bar{X} = A + \frac{\Sigma fd_x}{N} \times h = 150 + \frac{12}{66} \times 50 = 150 + 9.09 = 159.09$$

$$\bar{Y} = B + \frac{\Sigma fd_y}{N} \times k = 30 + \frac{-26}{66} \times 10 = 30 - 3.94 = 26.06$$

$$b_{xy} = \frac{N\Sigma fd_xd_y - \Sigma fd_x\Sigma fd_y}{N\Sigma fd_y^2 - (\Sigma fd_y)^2} \times \frac{h}{k} = \frac{66(-14) - 12(-26)}{66(100) - (-26)^2} \times \frac{50}{10}$$

$$= \frac{-924 + 312}{6600 - 676} \times \frac{50}{10} = -\frac{3060}{5924} = -0.5165.$$

Therefore, the regression equation of X on Y is : $X - 159.09 = -0.5165\,(Y - 26.06)$

$$X - 159.09 = -0.5165Y + 13.46 \text{ or } X = 172.55 - 0.5165Y$$

Regression equation of Y on X : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$b_{yx} = \frac{N\Sigma fd_xd_y - \Sigma fd_x\Sigma fd_y}{N\Sigma fd_x^2 - (\Sigma fd_x)^2} \times \frac{k}{h} = \frac{66(-14) - 12(-26)}{66(70) - (12)^2} \times \frac{10}{50} = -0.0273$$

Therefore, the regression equation of Y on X is : $Y - 26.06 = -0.0273\,(X - 159.09)$

$$Y = 26.06 - 0.0273X + 4.343 = 30.40 - 0.0273X.$$

The sales revenue corresponding to advertising expenditure of Rs. 50 thousand

(i)
$$X_{50} = 172.55 - 0.5165\,(50)$$
$$= 172.55 - 25.825 = 146.725$$

(*ii*) $Y_{300} = 30.40 - 0.0273\,(300)$
$= 30.40 - 8.19 = 22.21$

Hence, to attain sales revenue of Rs. 300 lakhs, the advertising expenditure required is Rs. 22.21 lakhs.

(*iii*) $r = \sqrt{b_{xy} \times b_{yx}}$

$= \sqrt{5165 \times .0273} = -0.119.$

Standard Error of Estimate

As we find it necessary to supplement an average with a measure of dispersion or variation, so in order to see how good or representative the regression line is, we look for a measure of variation about it. If we have a wide scatter or variation of the dots about the regression line, then it would have to be considered a poor representative of the relationship. The more closely the dots cluster around the line, the more representative it is and the better the estimate based on the equation for this line. And if the dots should all lie on the regression line a (hypothetical situation), then there is no variation about the line and the correlation is perfect.

The variation about the line of average relationship can be measured in the manner similar to the measuring of the variation of the items about an average. Thus, we use here a measuring of variation similar to the standard deviation—**the standard error of estimate.**

The measure of variation of the observations around the computed regression line is referred to as the standard error of estimate. Just as the standard deviation is a measure of the scatter of observations in a frequency distribution around the mean of that distribution, the standard error of estimate is a measure of the scatter of the observed values of Y around the corresponding computed values of Y on the regression line. It is computed as a standard deviation, being also a square root of the mean of the squared deviation. But the deviations here are not the deviations of the items from the arithmetic mean ; they are rather the vertical distances of every dot from the line of average relationship.

The deviation of each dot from the regression line is symbolised by $Y - Y_c$. Thus the square root of mean of the squared deviation is :

$$\sqrt{\frac{\Sigma (Y - Y_c)^2}{N - 2}}$$

This formula is not convenient from the computational point of view because it requires the computation of Y_c, *i.e.,* estimated values of Y. A more convenient formula is given below :

$$S_{y.x} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y + b\Sigma YX}{N - 2}}$$

where $S_{y.x}$ denote the S.E. of estimate of regression equation of y on x.

Similarly, we can calculate $S_{x.y}$.

$$S_{x.y} = \sqrt{\frac{\Sigma (X - X_c)^2}{N - 2}}$$

or

$$S_{x.y} = \sqrt{\frac{\Sigma X^2 - a\Sigma X + b\Sigma XY}{N - 2}}$$

The standard error of estimate can very easily be calculated with the help of the following formula :

$$S_{x.y} = S_y\sqrt{1-r^2} \; ; \qquad S_{y.x} = S_x\sqrt{1-r^2}$$

The standard error of estimate measures the accuracy of the estimated figures. The smaller the value of standard error of estimate, the closer will be dots to the regression line and the better the estimates based on the equation for this line. If standard error of estimate is zero, then there is no variation about the line and the correlation will be perfect. Thus with the help of standard error of estimate it is possible for us to ascertain how good and representative the regression line is as a description of the average relationship between two series.

Coefficient of Determination

The ratio of the unexplained variation to the total variation represents the proportion of variation in Y that is not explained by regression on X. Subtraction of this proportion from 1.0 gives the proportion of variation in Y that is explained by regression on X. The statistic used to express this proportion is called the coefficient of determination and is denoted by R^2. It may be written as follows :

$$R^2 = 1 - \frac{\text{Variation in } Y \text{ remaining after regression on } X}{\text{Total variation in } Y}$$

$$R^2 = 1 - \frac{\text{Error sum of squares}}{\text{Total sum of squares}}$$

The value of R^2 is the proportion of the variation in the dependent variable Y explained by regression on the independent variable X.

MISCELLANEOUS ILLUSTRATIONS

Illustration 8. Given the following bivariate data :

X :	–1	5	3	2	1	1	7	3
Y :	–6	1	0	0	1	2	1	5

(*a*) Fit a regression line of Y on X and predict Y if $X = 10$.

(*b*) Fit a regression line of X on Y and predict X if $Y = 2.5$. *(MBA, Osmania Univ., 2001)*

Solution.

FITTING REGRESSION EQUATIONS

X	$(X-3)$ d_x	d_x^2	Y	$(Y-2)$ d_y	d_y^2	$d_x d_y$
–1	–4	16	–6	–8	64	+32
5	+2	4	+1	–1	1	–2
3	0	0	0	–2	4	0
2	–1	1	0	–2	4	+2
1	–2	4	+1	–1	1	+2
1	–2	4	+2	0	0	–0
7	+4	16	+1	–1	1	–4
3	0	0	+5	+3	9	0
$\Sigma X = 21$	$\Sigma d_x = -3$	$\Sigma d_x^2 = 45$	$\Sigma Y = 4$	$\Sigma d_y = -12$	$\Sigma d_y^2 = 84$	$\Sigma d_x d_y = 30$

(a) $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{(8)(30) - (-3)(-12)}{(8)(45) - (-1)^2}$$

$$= \frac{240 - 36}{360 - 1} = \frac{204}{359} = 0.568$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{4}{8} = 0.5; \quad \bar{X} = \frac{\Sigma X}{N} = \frac{21}{8} = 2.625$$

$$Y - 0.5 = .568\,(X - 2.625)$$

$$Y = .568X - 0.991$$

(b) $X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$= \frac{(8)(30) - (-3)(-12)}{(8)(84) - (-12)^2} = \frac{204}{528} = 0.386$$

$$X - 2.625 = .386\,(Y - .5)$$

$$X = .386Y + 2.432$$

If $Y = 2.5$, X shall be

$$X = .386\,(2.5) + 2.432 = 3.397.$$

Illustration 9. From the following data obtain the two regression equations :

Sales	: 91	97	108	121	67	124	51	73	111	57
Purchase	: 71	75	69	97	70	91	39	61	80	47

Solution : CALCULATION OF REGRESSION EQUATIONS

Sales X	$(X - \bar{X})$ $\bar{X} = 90$ x	x^2	*Purchase* Y	$(Y - \bar{Y})$ $\bar{Y} = 70$ y	y^2	xy
91	+1	1	71	+1	1	+1
97	+7	49	75	+5	25	+35
108	+18	324	69	−1	1	−18
121	+31	961	97	+27	729	+837
67	−23	529	70	0	0	0
124	+34	1156	91	+21	441	+714
51	−39	1521	39	−31	961	+1209
73	−17	289	61	−9	81	+153
111	+21	441	80	+10	100	+210
57	−33	1089	47	−23	529	+759
$\Sigma X = 900$	$\Sigma x = 0$	$\Sigma x^2 = 6360$	$\Sigma Y = 700$	$\Sigma y = 0$	$\Sigma y^2 = 2868$	$\Sigma xy = 3900$

Regression equation of X on Y : $X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{900}{10} = 90; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{700}{10} = 70$$

$$b_{xy} = \frac{\Sigma XY}{\Sigma y^2} = \frac{3900}{2868} = 1.36$$

$$X - 90 = 1.36\,(Y - 70)$$

$$X - 90 = 1.36Y - 95.2 \text{ or } X = -5.2 + 1.36Y$$

Regression equation of Y on X : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{3900}{6360} = 0.613$$

$$Y - 70 = 0.613\,(X - 90)$$

$$Y - 70 = 0.613X - 55.17 \quad \text{or} \quad Y = 14.83 + 0.613\,X.$$

Illustration 10. The personnel manager of an electronic manufacturing company devises a manual dexterity test for job applicants to predict their production rating in the assembly department. In order to do this he selects a random sample of 10 applicants. They are given the test and later assigned a production rating. Results are as follows :

Worker	:	A	B	C	D	E	F	G	H	I	J
Test Score	:	53	36	88	84	86	64	45	48	39	69
Production Rating	:	45	43	89	79	84	66	49	48	43	76

Fit a linear least square regression equation of production rating on test score. *(MBA, Delhi Univ., 2002)*

Solution. Let test score be denoted by X and production rating by Y. We have to fit a regression equation of Y on X.

FITTING REGRESSION EQUATION OF Y ON X

Worker	X	$(X-61)$ d_x	d_x^2	Y	$(Y-62)$ d_y	$d_x d_y$
A	53	–8	64	45	–17	+136
B	36	–25	625	43	–19	+475
C	88	+27	729	89	+27	+729
D	84	+23	529	79	+17	+391
E	86	+25	625	84	+22	+550
F	64	+3	9	66	+4	+12
G	45	–16	256	49	–13	+208
H	48	–13	169	48	–14	+182
I	39	–22	484	43	–19	+418
J	69	+8	64	76	+14	+112
	$\Sigma X = 612$	$\Sigma d_x = 2$	$\Sigma d_x^2 = 3554$	$\Sigma Y = 622$	$\Sigma d_y = 2$	$\Sigma d_x d_y = 3213$

Regression Equation of Y on X : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10 \times 3213 - 2 \times 2}{10 \times 3554 - (2)^2} = \frac{32130 - 4}{35536} = \frac{32126}{35536} = +0.904$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{622}{10} = 62.2; \quad \bar{X} = \frac{\Sigma X}{N} = \frac{612}{10} = 61.2$$

Hence $Y - 62.2 = 0.904(X - 61.2)$

$$Y = .904X - 55.325 + 62.2$$

$Y = 6.875 + 0.904X$ is the required regression equation to predict production rating on test score.

Illustration 11. The following data give the ages and blood pressure of 10 women :

Age (X)	:	56	42	36	47	49	42	60	72	63	55
Blood Pressure (Y)	:	147	125	118	128	145	140	155	160	149	150

(*i*) Find the correlation coefficient between X and Y.

(*ii*) Determine the least square regression equation of Y on X.

(*iii*) Estimate the blood pressure of a woman whose age is 45 years.

Solution. CALCULATION OF CORRELATION COEFFICIENT

Age	*(X – 49)*		*Blood pressure*	*(Y – 145)*		
X	d_x	d_x^2	Y	d_y	d_y^2	d_xd_y
56	+7	49	147	+2	4	+14
42	–7	49	125	–20	400	+140
36	–13	169	118	–27	729	+351
47	–2	4	128	–17	289	+34
49	0	0	145	0	0	0
42	–7	49	140	–5	25	+35
60	+11	121	155	+10	100	+110
72	+23	529	160	+15	225	+345
63	+14	196	149	+4	16	+56
55	+6	36	150	+5	25	+30
$\Sigma X = 522$	$\Sigma d_x = 32$	$\Sigma d_x^2 = 1202$	$\Sigma Y = 1417$	$\Sigma d_y = -33$	$\Sigma d_y^2 = 1813$	$\Sigma d_xd_y = 1115$

(*i*) Coefficient of correlation is given by

$$r = \frac{N\Sigma d_xd_y - \Sigma d_x \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{10(1115) - (32)(-33)}{\sqrt{10(1202) - (32)^2}\sqrt{10(1813) - (-33)^2}}$$

$$= \frac{11150 + 1056}{\sqrt{12020 - 1024}\sqrt{18130 - 1089}} = \frac{12206}{13689} = 0.892$$

There is a high degree of positive correlation between age and blood pressure.

(*ii*) The least square regression equation of Y on X is given by

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{522}{10} = 52.2; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{1417}{10} = 141.7$$

and $$b_{yx} = \frac{N\Sigma d_xd_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10(1115) - 32(-33)}{10(1202) - (32)^2} = \frac{12206}{10996} = 1.11$$

Substituting these values in the above equation, we have

$$Y - 141.7 = 1.11(X - 52.2)$$

or $$Y = 1.11X + 141.7 - 57.942 = 83.758 + 1.11X$$

This is the required least square regression equation of Y on X.

(*iii*) When $X = 45$, then

$$Y = 83.758 + 1.11(45) = 83.758 + 49.95 = 133.708$$

Hence, the most likely blood pressure of a woman of 45 years is 134.

Illustration 12. For the following data determining to production and capacity utilisation :

	Average	*Standard deviation*
Production (in lakh units)	35.6	10.5
Capacity utilisation (in percentage)	84.8	8.5

$r = 0.62$

(*i*) Estimate the production when the capacity utilisation is 70 per cent.

(*ii*) The capacity utilisation to achieve the production of 50 lakh units.

(MBA, Pune Univ.; MBA, Delhi Univ., 2005)

Solution. Let production be denoted by the variable X and capacity utilisation by Y. Then the regression equation showing the regression equation of capacity utilisation of production will be given by the following formula :

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

where $$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.62 \times \frac{8.5}{10.5} = \frac{5.27}{10.5} = 0.5019$$

and $$\bar{X} = 35.6;\ \bar{Y} = 84.8$$

Substituting all these values in the above equation, we get

$$Y - 84.8 = 0.5019\,(X - 35.6)$$

or $$Y = 84.8 + 0.5019\,X - 17.8676 \text{ or } Y = 66.9324 + 0.5019X$$

which is the required regression of capacity utilisation on production.

To estimate the production, we shall have to find the regression equation of X on Y, *i.e.*,

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

where $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.62 \times \frac{10.5}{8.5} = \frac{6.51}{850} = 0.7659$$

Substituting the values, we have

$$(X - 35.6) = 0.7659\,(Y - 84.8)$$

or $$X = 35.6 + 0.7659\,Y - 64.9483$$

$$= -29.3483 + 0.7659\,Y$$

when $$Y = 70,\ X = -29.3483 + 0.7659\,(70)$$

$$= -29.3483 + 53.613 = 24.2647$$

Hence the estimated production is 2,42,647 units when the capacity utilisation is 70 per cent.

Illustration 13. There are two series of index numbers, P for price index and S for stock of a commodity. The mean and standard deviation of P are 100 and 8 and of S are 103 and 4 respectively. The correlation coefficient between the two series is 0.4. With these data, work out a linear equation to read off values of P for various values of S. Can the same equation be used to read off values of S for various values of P?

Solution. We have to fit an equation $P = a + bS$

$$(P - \bar{P}) = r\frac{\sigma_P}{\sigma_S}(S - \bar{S})$$

$$\bar{P} = 100,\ \bar{S} = 103,\ \sigma_P = 8,\ \sigma_S = 4,\ r = 0.4$$

$$\therefore \quad P - 100 = .4\,\frac{8}{4}(S - 103) \quad \text{or} \quad P = 17.6 + 0.8\,S.$$

The same equation cannot be used to read off values of S for various values of P. For that we have to fit an equation $S = a + bP$.

$$(S - \bar{S}) = r\frac{\sigma_S}{\sigma_P}(P - \bar{P})$$

$$S - 103 = 0.4\,\frac{4}{8}(P - 100) \text{ or } S = 83 + 0.2\,P.$$

Illustration 14. The following data show the experience of machine operators and their performance ratings as given by the number of good parts turned out per 100 pieces :

Operator	:	1	2	3	4	5	6	7	8
Experience (X)	:	16	12	18	4	3	10	5	12
Performance Rating (Y)	:	87	88	89	68	78	80	75	83

Calculate the regression line of performance ratings on experience and estimate the probable performance if an operator has 10 years' experience. *(MBA, Kumaun Univ., 1999)*

Solution. Let performance rating be denoted by Y and experience by X. We have to calculate the regression line of Y on X.

CALCULATION OF REGRESSION EQUATIONS

Experience X	$(X-10)$ x	x^2	*Performance* Y	$(Y-81)$ y	y^2	xy
16	+6	36	87	+6	36	+36
12	+2	4	88	+7	49	+14
18	+8	64	89	+8	64	+64
4	−6	36	68	−13	169	+78
3	−7	49	78	−3	9	+21
10	0	0	80	−1	1	0
5	−5	25	75	−6	36	+30
12	+2	4	83	+2	4	+4
$\Sigma X = 80$	$\Sigma x = 0$	$\Sigma x^2 = 218$	$\Sigma Y = 648$	$\Sigma y = 0$	$\Sigma y^2 = 368$	$\Sigma xy = 247$

Regression equation of Y on X : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{247}{218} = 1.133; \quad \overline{Y} = \frac{648}{8} = 81, \quad \overline{X} = \frac{80}{8} = 10$$

$\therefore$ $Y - 81 = 1.133\,(X - 10) = 1.133\,X - 11.33$

$Y = 69.67 + 1.133X$

When $X = 10$, Y will be

$Y = 69.67 + 1.133\,(10) = 69.67 + 11.33 = 81$

Thus the probable performance of an operator who has 10 years' experience is 81 good parts out of 100.

Illustration 15. Find the most likely production corresponding to a rainfall of 40" from the following data :

	Rainfall	*Production*
Average	30"	50 quintals
S.D.	5"	10 quintals
Coefficient of correlation	0.8	

Solution. Let rainfall be denoted by X and production by Y. The expected yield corresponding to a rainfall 40" will be obtained by the regression equation of Y on X.

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$\overline{Y} = 50$, $\sigma_y = 10$, $\overline{X} = 30$, $\sigma_x = 5$, $r = 0.8$

$$Y - 50 = 0.8\,\frac{10}{5}\,(X - 30) \text{ or } Y - 50 = 1.6\,(X - 30)$$

$Y - 50 = 1.6X - 48$ or $Y = 2 + 1.6X$

When rainfall (X) is 40", the expected production, *i.e.*, Y would be

$Y = 2 + 1.6\,(40) = 66$ quintals.

Illustration 16. Obtain the regression equations from the data given below :

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

Plot the regression equation on a graph paper and determine $\overline{X}$ and $\overline{Y}$. Also calculate the value of correlation coefficient.

(BBA, BHU, 2000; MBA, Hyderabad Univ., 2005)

Solution.

CALCULATION OF REGRESSION EQUATIONS

X	$(X - \overline{X})$ x	x^2	Y	$(Y - \overline{Y})$ y	y^2	xy
1	−4	16	9	−3	9	+12
2	−3	9	8	−4	16	+12
3	−2	4	10	−2	4	+ 4
4	−1	1	12	0	0	0
5	0	0	11	−1	1	0
6	+1	1	13	+1	1	+1
7	+2	4	14	+2	4	+4
8	+3	9	16	+4	16	+12
9	+4	16	15	+3	9	+12
$\Sigma X = 45$	$\Sigma x = 0$	$\Sigma x^2 = 60$	$\Sigma Y = 108$	$\Sigma y = 0$	$\Sigma y^2 = 60$	$\Sigma xy = 57$

Regression Equation of Y on X : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{45}{9} = 5;\ \bar{Y} = \frac{\Sigma Y}{N} = \frac{108}{9} = 12$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{57}{60} = 0.95$$

$Y - 12 = 0.95\,(X - 5) = 0.95\,X - 4.75$ or $Y = 7.25 + 0.95\,X$

Regression Equation of X on Y : $X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{57}{60} = 0.95$$

$X - 5 = 0.95\,(Y - 12) = 0.95Y - 11.4$ or $X = -6.4 + 0.95Y$

Graphing regression equations

From the regression equation of Y on X, we can estimate the most probable value of Y for various values of X and from the regression equation of X on Y, we can estimate the most probable values of X for various values of Y.

		(*Estimated value of Y*)
		$Y = 7.25 + 0.95X$
when	$X = 1$,	$Y = 7.25 + .95(1) = 8.20$
when	$X = 2$,	$Y = 7.25 + .95(2) = 9.15$
when	$X = 3$,	$Y = 7.25 + .95(3) = 10.10$
when	$X = 4$,	$Y = 7.25 + .95(4) = 11.05$
when	$X = 5$,	$Y = 7.25 + .95(5) = 12.00$
when	$X = 6$,	$Y = 7.25 + .95(6) = 12.95$
when	$X = 7$,	$Y = 7.25 + .95(7) = 13.90$
when	$X = 8$,	$Y = 7.25 + .95(8) = 14.85$
when	$X = 9$,	$Y = 7.25 + .95(9) = 15.80$

To plot the regression line of Y on X, we will take the actual values of X and estimated values of Y.

		(*Estimated values of X*)
		$X = -6.4 + 0.95Y$
when	$Y = 9$,	$X = .95(9) - 6.4 = 2.15$
when	$Y = 8$,	$X = .95(8) - 6.4 = 1.20$
when	$Y = 10$,	$X = .95(10) - 6.4 = 3.10$
when	$Y = 11$,	$X = .95(11) - 6.4 = 4.05$
when	$Y = 12$,	$X = .95(12) - 6.4 = 5.00$
when	$Y = 13$,	$X = .95(13) - 6.4 = 5.95$
when	$Y = 14$,	$X = .95(14) - 6.4 = 6.90$
when	$Y = 15$,	$X = .95(15) - 6.4 = 7.85$
when	$Y = 16$,	$X = .95(16) - 6.4 = 8.80$

Coefficient of Correlation :

We are given $b_{xy} = 0.95$ and $b_{yx} = 0.95$

$r = \sqrt{b_{yx} \times b_{xy}}$ or $r = \sqrt{.95 \times .95} = 0.95$

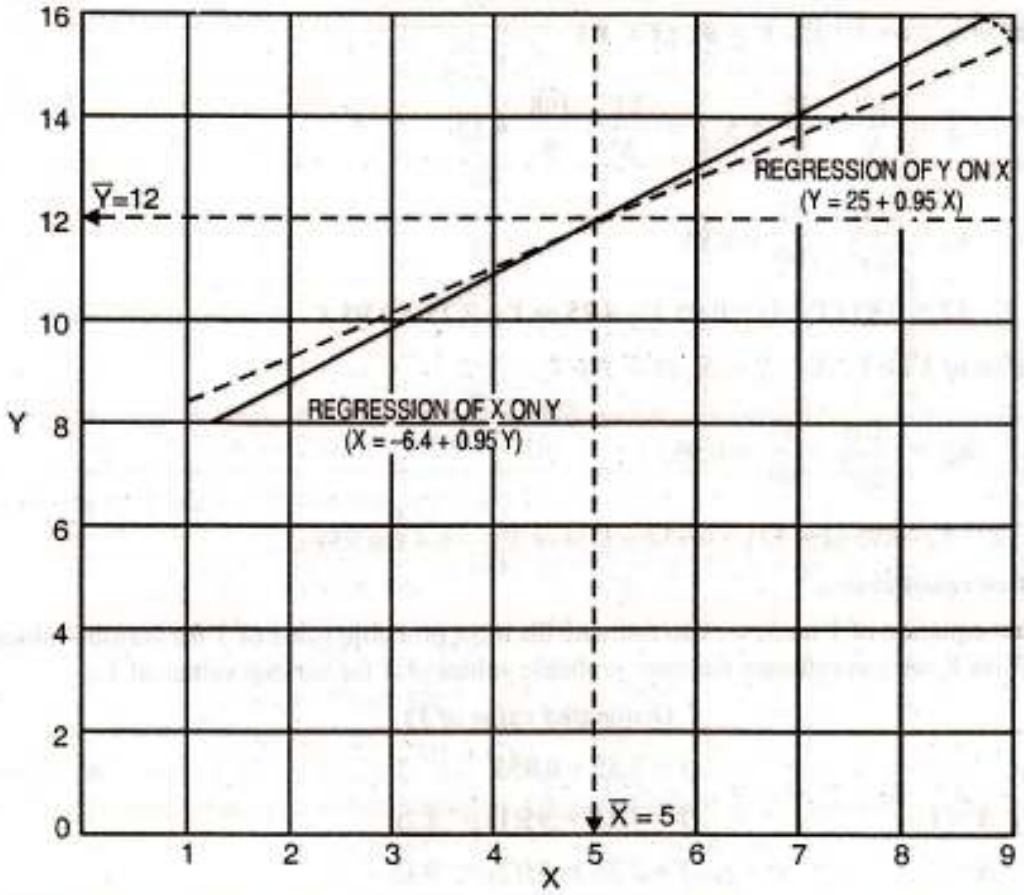


Illustration 17. The General Sales Manager of Kiran Enterprises—an enterprise dealing in the sale of ready-made men's wears—is toying with the idea of increasing his sales to Rs. 80,000. On checking the records of sales during the last 10 years, it was found that the annual sale proceeds and advertisement expenditure were highly correlated to the extent of 0.8. It was further noted that the annual average sale has been Rs. 45,000 and annual average advertisement expenditure Rs. 30,000, with a variance of Rs. 1,600 and Rs. 626 in advertisement expenditure respectively.

In view of the above, how much expenditure on advertisement you would suggest the General Sales Manager of the enterprise to incur to meet his target of sales. ***(MBA, Kurukshetra Univ., 2004)***

Solution. Let advertisement expenditure be denoted by X and sales by Y. We are required to find out the regression equation of Y on X given by the equation,

$$(Y - \bar{Y}) = r \frac{\sigma_y}{\sigma_x} (X - \bar{X}).$$

$$r = 0.8, \quad \sigma_x = 400, \quad \sigma_y = 25, \quad \bar{X} = 45{,}000, \quad \bar{Y} = 30{,}000$$

Substituting the values, we get

$$(Y - 30{,}000) = 0.8 \frac{25}{400} (X - 45{,}000) = .05 (X - 45{,}000)$$

$$Y = 30{,}000 + .05X - 2{,}250 = 27{,}750 + .05X$$

When $X = 80{,}000$

$$Y = 27{,}750 + .05 \times 80{,}000 = 27{,}750 + 4{,}000 = 31{,}750$$

Hence the General Sales Manager should spend Rs. 31,750 to have the target sales of Rs. 80,000.

Illustration 18. Suppose that you are interested in using past expenditure on research and development by a firm to predict current expenditures on R & D. You got the following data by taking a random sample of firms, where X is the amount on R & D (in lakhs of rupees) 5 years ago and Y is the amount spent on R & D (in lakhs of rupees) in the current year :

X	:	30	50	20	80	10	20	20	40
Y	:	50	80	30	110	20	20	40	50

(*i*) Find the regression equation of Y on X.

(*ii*) If a firm is chosen randomly and $X = 10$, can you use the regression to predict the value of Y? Discuss.

(MBA, Madurai-Kamaraj Univ., 2000)

Solution. CALCULATION OF REGRESSION EQUATION

X	$(X-33)$ d_x	d_x^2	Y	$(Y-50)$ d_y	d_y^2	d_xd_y
30	–3	9	50	0	0	0
50	+17	289	80	+30	900	+510
20	–13	169	30	–20	400	+260
80	+47	2209	110	+60	3600	+2820
10	–23	529	20	–30	900	+690
20	–13	169	20	–30	900	+390
20	–13	169	40	–10	100	+130
40	+7	49	50	0	0	0
$\Sigma X = 270$	$\Sigma d_x = +6$	$\Sigma d_x^2 = 3592$	$\Sigma Y = 400$	$\Sigma d_y = 0$	$\Sigma d_y^2 = 6800$	$\Sigma d_xd_y = 4800$

(*i*) *Regression equation of Y on X* : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{270}{8} = 33.75; \quad \overline{Y} = \frac{\Sigma Y}{N} = \frac{400}{8} = 50$$

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8 \times 4800 - 6 \times 0}{8 \times 3592 - (6)^2} = \frac{38400}{28700} = 1.338$$

$$Y - 50 = 1.338\,(X - 33.75) \text{ or } Y = 4.84 + 1.338X$$

(*ii*) When X is 10 : $Y = 4.84 + 1.338\,(10) = 18.22$

For $X = 10$, Y is 18.22.

Illustration 19. You are given the following information about advertising expenditure and sales :

	Adv. Exp. (X) *(Rs. lakhs)*	*Sales* (Y) *(Rs. lakhs)*
$\overline{X}$	10	90
σ	3	12

Correlation coefficient = 0.8

(*i*) Obtain the two regression equations.

(*ii*) Find the likely sales when advertisement budget is Rs. 15 lakhs.

(*iii*) What should be the advertisement budget if the company wants to attain sales target of Rs. 120 lakhs?

(MBA, Kumaun Univ., 2000; MBA, DU, 2002, MBA (HCA), DU, 2003)

Solution. (*i*) *Regression equation of X on Y* : $X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$

$$\overline{X} = 10,\ r = 0.8,\ \sigma_x = 3,\ \sigma_y = 12,\ \overline{Y} = 90$$

$$X - 10 = 0.8\,\frac{3}{12}\,(Y - 90)$$

$$X - 10 = 0.2\,(Y - 90)$$

$$X - 10 = 0.2Y - 18 \quad \text{or} \quad X = -8 + 0.2Y$$

Regression equation of Y on X : $Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$

$$Y - 90 = .8\,\frac{12}{3}\,(X - 10)$$

$$Y - 90 = 3.2\,(X - 10) \text{ or } Y = 58 + 3.2X$$

(*ii*) By putting 15 in regression equation of Y on X, we can find out the likely sales,

$$Y = 58 + 3.2\,(15) = 58 + 48 = 106$$

Thus the likely sales for advertisement budget of Rs. 15 lakhs is Rs. 106 lakhs.

(*iii*) By putting 120 in regression equation of X on Y, we can find what should be the advertisement budget.

$$X = -8 + 0.2\,(120) = 16$$

Thus for attaining sales target of Rs. 120 lakhs, the advertisement budget should be Rs. 16 lakhs.

Illustration 20. The following table gives the aptitude test scores and productivity indices of 10 workers selected at random :

Aptitude scores	:	60	62	65	70	72	48	53	73	65	82
Productivity index	:	68	60	62	80	85	40	52	62	60	81

Estimate (*i*) the productivity index of a worker whose test score is 92, (*ii*) the test score of a worker whose productivity index is 75. *(MBA, Delhi Univ., 2001; MBA, Hyderabad, Univ., 2004)*

Solution. Since productivity depends on aptitude scores, let Y denote the productivity and X the aptitude score.

CALCULATION OF REGRESSION EQUATIONS

Aptitude Score X	$(X-65)$ $\bar{X}=65$ x	x^2	*Productivity Index* Y	$(Y-65)$ $\bar{Y}=65$ y	y^2	xy
60	–5	25	68	+3	9	–15
62	–3	9	60	–5	25	+15
65	0	0	62	–3	9	0
70	+5	25	80	+15	225	+75
72	+7	49	85	+20	400	+140
48	–17	289	40	–25	625	+425
53	–12	144	52	–13	169	+156
73	+8	64	62	–3	9	–24
65	0	0	60	–5	25	0
82	+17	289	81	+16	256	+272
$\Sigma X = 650$	$\Sigma x = 0$	$\Sigma x^2 = 894$	$\Sigma Y = 650$	$\Sigma y = 0$	$\Sigma y^2 = 1752$	$\Sigma xy = 1044$

For answering part (*i*) of the question we have to fit a regression equation of Y on X.

$$Y - \bar{Y} = b_{yx}\,(X - \bar{X})$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{650}{10} = 65;\quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{650}{10} = 65$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{1044}{894} = 1.168$$

$$Y - 65 = 1.168\,(X - 65)$$

$$Y - 65 = 1.168\,X - 75.92 \quad \text{or} \quad Y = 1.168\,X - 10.92$$

$$Y_{92} = 1.168\,(92) - 10.92 = 107.456 - 10.92 = 96.536$$

For answering part (*ii*) of the question we have to fit a regression equation of X on Y.

$$X - \bar{X} = b_{xy}\,(Y - \bar{Y})$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{1044}{1752} = 0.596$$

$$X - 65 = 0.596\,(Y - 65) \quad \text{or} \quad X - 65 = 0.596Y - 38.74$$

$$X = 0.596Y + 26.26$$

$$Y_{75} = .596\,(75) + 26.26 = 44.7 + 26.26 = 70.96.$$

Illustration 21. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible :

Variance of $X = 9$

Regression equation $8X - 10Y + 66 = 0$

$40X - 18Y = 214$

Find on the basis of the above information :

(*i*) The mean values of X and Y,

(*ii*) Coefficient of correlation between X and Y, and

(*iii*) Standard deviation of Y. *(MBA, Pune Univ., 2002; MBA, Anna Univ., 2003)*

Solution. (*i*) *Calculating mean values of X and Y*

$$8X - 10Y = -66 \quad ...(i)$$

$$40X - 18Y = 214 \quad ...(ii)$$

Multiplying eq. (*i*) by 5

$$40X - 50Y = -330$$

$$\begin{array}{r} 40X - 18Y = 214 \\ - \quad + \quad - \\ \hline -32Y = -544 \end{array}$$

$$Y = 17 \quad \text{or} \quad \overline{Y} = 17$$

Putting the value of Y in eq. (*i*)

$$8X - 10\,(17) = -66$$

$$8X = -66 + 170$$

$$8X = 104 \text{ or } X = 13 \text{ or } \overline{X} = 13$$

(*ii*) *Coefficient of correlation between X and Y*

For finding the value of r, we have to determine the value of regression coefficients. Since we don't know which equation is regression of X on Y and which is of Y on X, we have to make an assumption. Assuming eq. (*i*) as the regression of X on Y.

$$8X = 10Y - 66$$

$$X = -\frac{66}{8} + \frac{10}{8}Y \quad \text{or } b_{xy} = \frac{10}{8}$$

From eq. (*ii*) $-18Y = 214 - 40X$

$$Y = -\frac{214}{18} + \frac{40}{18}X \text{ or } b_{yx} = \frac{40}{18}$$

Since both the regression coefficients are greater than 1, our assumption is wrong. Hence eq. (*i*) is regression eq. of Y on X.

$$-10\,Y = -66 - 8X$$

$$Y = \frac{66}{10} + \frac{8}{10}X \quad or \quad b_{yx} = \frac{8}{10}$$

From eq. (*ii*) $40X = 214 + 18Y$

$$X = \frac{214}{40} + \frac{18}{40}Y \quad or \quad b_{xy} = \frac{18}{40}$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{18}{40} \times \frac{8}{10}} = \sqrt{0.36} = 0.6$$

(*iii*) The value of standard deviation of Y can be determined from any regression coefficient.

$$b_{xy} = r\,\frac{\sigma_x}{\sigma_y}$$

$$b_{xy} = \frac{18}{40}, \; r = .6, \; \sigma_x = \sqrt{9} = 3$$

Substituting the values

$$\frac{18}{40} = .6\frac{3}{\sigma_y} \quad \text{or} \quad 18\sigma_y = 72 \text{ or } \sigma_y = 4.$$

Illustration 22. The coefficient of correlation between the ages of husbands and wives in a community was found to be +0.8, the average of husbands age was 25 years and that of wives age 22 years. Their standard deviations were 4 and 5 years respectively. Find with the help of regression equations :

(*a*) the expected age of husband when wife's age is 16 years, and

(*b*) the expected age of wife when husband's age is 33 years. *(MBA, Osmania Univ., 2000)*

Solution. Let age of wife be denoted by Y and age of husband by X. We are given

$$\overline{X} = 25,\ \overline{Y} = 22,\ \sigma_x = 4,\ \sigma_y = 5,\ r = 0.8$$

For answering part (*a*) we have to fit a regression equation X on Y

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

$$X - 25 = .8\frac{4}{5}(Y - 22) \text{ or } X - 25 = .64(Y - 22)$$

$$X - 25 = .64Y - 14.08 \text{ or } X = 10.92 + 0.64Y$$

When $Y = 16$, $X = 10.92 + 0.64\,(16) = 10.92 + 10.24 = 21.16$

Thus, the expected age of husband when wife's age is 16 years shall be 21.16 years.

For answering part (*b*) we have to fit a regression equation of Y on X.

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$Y - 22 = .8\,\frac{5}{4}(X - 25)$$

$$Y - 22 = (X - 25) \text{ or } Y = -3 + X; \text{ when } X = 33,$$

$$Y = -3 + 33 = 30$$

Thus, the expected age of wife when husband's age is 33 is 30 years.

Illustration 23. The following data relate to marks obtained by 250 students in Accountancy and Statistics in an examination of a university :

Subject	*Arithmetic Mean*	*Standard Deviation*
Accountancy	48	4
Statistics	55	5

Coefficient of correlation between marks in accountancy and statistics is +0.8. Find the two regression equations and estimate the marks obtained by a student in Statistics who secured 50 marks in Accountancy.

(M.Com., Sukhadia Univ., 2001)

Solution. Let marks in accountancy be denoted by X and in statistics by Y.

Regression equation of X on Y

$$X - \overline{X} = r\,\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

$$\overline{X} = 48,\ \overline{Y} = 55,\ \sigma_x = 4,\ \sigma_y = 5,\ r = 0.8$$

$$X - 48 = .8\,\frac{4}{5}(Y - 55)$$

$$X - 48 = .64\,(Y - 55)$$

$$X = .64Y + 12.8$$

Regression equation of Y on X

$$Y - \overline{Y} = r\,\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$Y - 55 = .8\,\frac{5}{4}(X - 48)$$

$$Y - 55 = (X - 48) \text{ or } Y = 7 + X$$

If marks in accountancy, *i.e.*, X is 50; the marks in statistics shall be 57.

Illustration 24. The following figures relate to length of service and income of the employees of an organisation :

Length of Service (Years) :	11	7	2	5	8	6	10
Income (Rs. hundred) :	7	5	3	2	6	4	8

Compute the coefficient of correlation for the above data. Find the two regression equations and examine the relationship.

Solution. Let length of service be denoted by X and income by Y.

CALCULATION OF REGRESSION EQUATIONS AND CORRELATION COEFFICIENT

X	$(X-7)$ x	x^2	Y	$(Y-5)$ y	y^2	xy
11	+4	16	7	+2	4	+8
7	0	0	5	0	0	0
2	−5	25	3	−2	4	+10
5	−2	4	2	−3	9	+6
8	+1	1	6	+1	1	+1
6	−1	1	4	−1	1	+1
10	+3	9	8	+3	9	+9
$\Sigma X = 49$	$\Sigma x = 0$	$\Sigma x^2 = 56$	$\Sigma Y = 35$	$\Sigma y = 0$	$\Sigma y^2 = 28$	$\Sigma xy = 35$

Regression equation of X on Y : $X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{49}{7} = 7; \quad \bar{Y} = \frac{35}{7} = 5$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{35}{28} = 1.25$$

$$X - 7 = 1.25\,(Y - 5)$$

$$X - 7 = 1.25Y - 6.25 \text{ or } X = 0.75 + 1.25Y$$

Regression equation of Y on X : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$b_{yx} = \frac{\Sigma xy}{\Sigma y^2} = \frac{35}{56} = 0.625$$

$$X - 5 = .625\,(X - 7)$$

$$X - 5 = .625X - 4.375 \text{ or } X = 0.625 + 0.625\,X$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{1.25 \times .625} = 0.884$$

Thus, there is a high degree of positive correlation between length of service and experience.

Illustration 25. In a correlation study the following values are obtained :

	X	Y
Mean	65	67
S.D.	2.5	3.5

Coefficient of Correlation $r = 0.8$.

Find the two regression equations. *(M.Com., Madurai-Kamaraj Univ., 2007)*

Solution : *Regression equation of X on Y :*

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$\bar{X} = 65, \sigma_x = 2.5, \sigma_y = 3.5, r = 0.8, \bar{Y} = 67$$

$$X - 65 = 0.8\,\frac{2.5}{3.5}\,(Y - 67)$$

$$X - 65 = 0.571\,(Y - 67)$$

$$X - 65 = 0.571\,Y - 38.26$$

$$X = 0.571Y + 26.74$$

Regression equation of Y on X :

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$Y - 67 = 0.8\,\frac{3.5}{2.5}\,(X - 65)$$

$$Y - 67 = 1.12\,(X - 65)$$

$$Y - 67 = 1.12\,X - 72.8$$

$$Y = 1.12\,X - 5.8$$

Illustration 26. In trying to evaluate the effectiveness in its advertising campaign, a firm compiled the following information :

Year	:	2003	2004	2005	2006	2007	2008	2009	2010
Adv. Expenditure ('000 Rs.)	:	12	15	15	23	24	38	42	48
Sales (Rs. lakh)	:	5.0	5.6	5.8	7.0	7.2	8.8	9.2	9.5

Calculate the regression equation of sales on advertising expenditure. Estimate the probable sales when advertisement expenditure is Rs. 60 thousand.

Solution : CALCULATION OF REGRESSION EQUATION

X	$(X-24)$ d_x	d_x^2	Y	$(Y-7.0)$ d_y	d_y^2	$d_x d_y$
12	−12	144	5.0	−2.0	4.00	24.0
15	−9	81	5.6	−1.4	1.96	12.6
15	−9	81	5.8	−1.2	1.44	10.8
23	−1	1	7.0	0	0	0
24	0	0	7.2	+ 0.2	.04	0
38	+14	196	8.8	+1.8	3.24	25.2
42	+18	324	9.2	+ 2.2	4.84	39.6
48	+24	576	9.5	+2.5	6.25	60.0
$\Sigma X = 217$	$\Sigma d_x = 25$	$\Sigma d_x^2 = 1403$	$\Sigma Y = 58.1$	$\Sigma d_y = 2.1$	$\Sigma d_y^2 = 21.77$	$\Sigma d_x d_y = 172.2$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{217}{8} = 27.125; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{58.1}{8} = 7.26$$

Regression equation of sales on advertisement expenditure is given by :

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X})$$

where

$$b_{yx} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{8(172.2) - (25)(2.1)}{8(1403) - (25)^2} = \frac{1377.6 - 52.5}{11224 - 625} = \frac{1325.1}{10599}$$

$$= 0.125$$

Substituting the values, we have

$$Y - 7.2625 = 0.125\,(X - 27.125)$$
$$Y - 7.2625 = 0.125X - 3.3906$$
$$Y = 3.8719 + 0.1250X$$

When $X = 60$, the estimated value of Y shall be :

$$Y = 3.8719 + 0.1250\,(60) = 3.8719 + 7.5 \approx 11.37.$$

Illustration 27. A resarch company summarized advertising expenditure and sales results as follows :

	Ad. Expenditure (Rs. crore)	*Sales (Rs. crore)*
Mean	20	200
S.D.	18	170
r	$= 0.6.$	

Derive two regression equations. *(MBA, GGDIP Univ., 2009)*

Solution : Since sales depend on advertisement expenditure, we take sales as Y and advertisement expenditure as X.

Regression equation of X on Y :

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$\bar{X} = 20, \sigma_x = 18, \sigma_y = 170, r = 0.6, \bar{Y} = 200$$

$$X - 20 = 0.6\,\frac{18}{170}\,(Y - 200)$$
$$X - 20 = 0.64\,(Y - 200)$$
$$X - 20 = 0.64\,Y - 12.8$$
$$X = 0.64Y + 7.2$$

Regression equation of Y on X :

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

$$\bar{Y} = 200, \sigma_y = 170, \sigma_x = 18, r = 0.6, \bar{X} = 20$$

$$Y - 200 = 0.6 \frac{170}{18} (X - 20)$$

$$Y - 200 = 5.667 (X - 20)$$

$$Y - 200 = 5.667 X - 113.34$$

$$Y = 5.667 X - 86.66$$

PROBLEMS

: Answer the following questions, each question carries **one** mark:

(*i*) What is regression ?

(*ii*) What is the use of studying regression ?

(*iii*) When will regression coefficients become coefficient of correlation ? *(MBA, Madurai-Kamaraj Univ., 2003)*

(*iv*) Write down the two regression equations.

(*v*) Write down the formula for regression coefficient of x and y ?

(*vi*) What do you understand by the term 'regression line' ? *(M.Com., M.K.Univ., 2003)*

(*vii*) What are regression coefficients ?

(*viii*) Can both the regression coefficients exceed one ?

(*ix*) Are regression coefficients independent of change of scale and origin or only origin ?

(*x*) In the regression equation of y on x how do you interpret the values of 'a' and 'b'?

(*xi*) Who had coined the term 'regression' ?

: Answer the following questions, each question carries **four** marks:

(*i*) Distinguish between 'correlation' and 'regression analysis'. Why there are two regression lines? *(MBA, UP Tech. Univ., 2007)*

(*ii*) What are regression coefficients ? How do you interpret them ?

(*iii*) What are the important characteristics of regression coefficients ?

(*iv*) If two regression coefficients are –1.2 and – 0.8, what would be the value of r ?

(*v*) What are the important uses of regression analysis ?

(*a*) Explain the concept of regression and point out its usefulness in dealing with business problems.

(*b*) Distinguish between correlation and regression. Also point out the properties of regression coefficients.

(*a*) Compare and contrast the role of correlation and regression in studying the interdependence of two variates.

(*b*) Explain the concept of regression and point out its importance in business forecasting.

Under what conditions can there be one regression line? Explain.

"The regression line gives only the best estimate of the value of quantity in question. We may assess the degree of uncertainty in this estimate by calculating a quantity known as the standard error of estimate". Elucidate.

Do you agree with the view that regression equations are irreversible, *i.e.*, we cannot find out the regression of X on Y from that of Y on X?

(*a*) Point out the usefulness of regression analysis in business and industry.

(*b*) What is linear regression? When is it used? *(MBA, Madurai-Kamaraj Univ., 2003)*

(*c*) Discuss the role of correlation and regression analysis in business. Illustrate.

What are regression lines ? With the help of an example, illustrate how they help in business decision-making. *(MBA, Delhi Univ., 2004)*

What do you understand by the term "regression analysis"? Point out the role of regression analysis in business decision-making. What are the important properties of regression coefficients? *(MBA, Osmania Univ.; MBA, Delhi Univ., 2006)*

(*a*) Write any two differences between correlation and regression. *(M.Com., Madras univ., 2009)*

(*b*) What are regression coefficients? State some of the important properties of regression coefficients.

(*c*) Write down the mathematical properties of Correlation Coefficient and Regression Coefficient. *(MBA, Hyderabad Univ., 2005)*

(*d*) State the utility of regression in economic analysis.

The following data give the hardness (X) and tensile strength (Y) of 7 samples of metal in certain units. Find the linear regression equation of Y on X.

X :	146	152	158	164	170	176	182
Y :	75	78	77	89	82	85	86

[$Y = 29.45 + 0.31X$]

12. The average daily wage for working class in Nagpur is Rs. 12 and for that in Delhi Rs. 18, their respective standard deviations are Rs. 2 and Rs. 3 and the coefficient of correlation is 0.67. Find the most likely wage in Delhi corresponding to the wage of Rs. 20 in Nagpur.

[$Y_{20} = 26.04$]

13. There are two series of index numbers D for disposable personal income and S for a salary of the company. The mean and standard deviations of the D series are 120 and 15 respectively and of the S series 115 and 10. The coefficient of correlation between the two series is 0.75. From the given information obtain a linear equation for estimating the values of S for different values of D. How will you interpret the values of S corresponding to different values of D obtained from the equation? Can the same equation be used for estimating values of D for different values of S?

[$S = 0.5$; $D = 55$; No]

14. The following calculations have been made for closing prices of 12 stocks (X) on the Bombay Stock Exchange on a certain day along with the volume of sales in thousand of shares (Y). From these calculations find the regression equations.

$\Sigma X = 580$, $\Sigma Y = 370$, $\Sigma XY = 11{,}494$

$\Sigma X^2 = 41{,}658$, $\Sigma Y^2 = 17{,}206$

[$Y = 53.55 - 0.47X$, $X = 79.16 - 1.1Y$]

15. Given the following data, what will be the possible yield when the rainfall is 29" ?

	Rainfall	*Production*
Mean	29"	40 units per acre
S.D.	3"	6 units per acre

Coefficient of correlation between rainfall and production = 0.8.

[40 units]

16. In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales:

Salesmen :	1	2	3	4	5	6	7	8	9	10
Test Scores :	45	75	50	60	80	90	85	40	80	55
Sales ('000) :	2.0	6.5	3.5	5.0	4.5	6.0	6.5	2.5	5.5	4.5

Calculate the regression line of sales on test score and estimate the most probable weekly sales volume if a salesman makes a score of 70.

[$Y = -0.541 + 0.078X$, 4919]

17. The following marks have been obtained by a group of students in Statistics (out of 100):

Paper I :	80	45	55	56	58	60	65	68	70	75	85
Paper II :	82	56	50	48	60	62	64	65	70	74	90

Compute the coefficient of correlation for the above data. Find the lines of regression and examine the relationship.

[$r = 0.75$, $Y = -1 + 0.75X$, $X = 4.25 + 0.75Y$]

18. The following table gives marks out of 50 awarded in a French and a German test to the same group of boys. Assume there is a linear relation between the sets of marks, calculate the equations of the lines of regression.

French :	10	10	18	25	28	33	34	39	42	43
German :	11	22	22	19	35	27	33	40	42	47

[$Y = 6.25 + 0.13X$, $X = -0.34 + 0.96Y$]

19. You are given the following result of the height (X) and weight (Y) of 1,000 managers:

Mean (X)	= 68.00"
Mean (Y)	= 150 lbs
Standard deviation (X)	= 2.50"
Standard deviation (Y)	= 20 lbs

Coefficient of correlation between X and Y = 0.6. Estimate from the above data the height of a manager whose weight is 200 lbs. *(MBA, Kurukshetra Univ., 2002)*

20. The following table shows the mean and standard deviation of the prices of two shares on a stock exchange :

Shares	*Mean (in Rs.)*	*Standard deviation (in Rs.)*
A Ltd.	39.5	10.8
B Ltd.	47.5	16.8

If the coefficient of correlation between the prices of two shares is 0.42, find the most likely price of share A corresponding to a price of Rs. 55 observed in the case of share B.

21. Catalogues listing textbooks were examined to discover the relationship between the cost of a book and number of pages it contains. The perusal gives the following data for ten books:

Pages	700	540	210	625	380	910	610	420	750	400
Price (Rs.) :	12	11	5	10	7	15	9	8	12	9

(*a*) Obtain the line of regression for estimating the price of a book.
(*b*) What is your estimate for the price of a book containing 500 pages ?
(*c*) What increase would you expect for a book if it is decided to increase the number of pages of the book by 100 ?
(*d*) Calculate the standard error of the estimate.

22. From the data given below find the two regression equations.

Age of wife	*Age of Husband*			
	20–25	*25–30*	*30–35*	*Total*
16–20	4	9	—	13
20–24	1	4	1	6
24–28	4	4	3	11
Total	9	17	4	30

(M.Phil, Kurukshetra Univ., 2003)

23. The data given below relate to the scores obtained by 9 salesmen in an intelligence test and their weekly sales, in lakh of rupees :

Salesman :	1	2	3	4	5	6	7	8	9
Test Score :	50	60	50	60	80	50	80	40	70
Sales (Rs. lakh) :	3	6	4	5	6	3	7	5	6

Obtain regression equation of sales on the intelligence test scores. If a salesman has obtained a score of 65, what would be his expected weekly sales ?
[$Y = 0.075X + 0.5$, Rs. 5.375 lakh]

24. The following figures relate to advertisement expenditure and sales :

Adv. Exp. (in lakh of Rs.) :	60	62	65	70	73	75	71
Sales (in crore of Rs.) :	10	11	13	15	16	19	14

Estimate (*i*) the sales for advertisement expenditure of Rs. 80 lakh and (*ii*) the advertisement for a sales target of Rs. 25 crore.
[20.1; 87.75]

25. You are given the following data about the sales and advertisement expenditure of a firm :

	Sales (Rs. crore)	*Advertisement Expenditure (Rs. crore)*
Arithmetic Mean	50	10
Standard Deviation	10	2
Coefficient of Correlation	+0.9	

(*a*) Calculate the two regression equations.
(*b*) Estimate the likely sales for a proposed advertisement expenditure of Rs. 13.5 crore.
(*c*) What should be the advertisement budget if the company wants to achieve a sales target of Rs. 70 crore ?

(MBA, Delhi Univ., 2005)

[(*a*) $Y = 4.5X + 5$, $X = .18Y + 1$. (*b*) 65.75 crore. (*c*) 13.6 crore]

26. The following bivariate frequency distribution relates to sales turnover (in lakh Rs.) and money spent on advertising budget (in thousand Rs.). Obtain the two regression equations.

Sales Turnover (in lakh Rs.)	*Advertising budget (in thousand Rs.)*			
	50–60	*60–70*	*70–80*	*80–90*
25–50	2	1	2	5
50–75	3	4	7	6
75–100	1	5	8	6
100–125	2	7	9	2

Estimate (*i*) the sales turnover corresponding to advertising budget of Rs. 150 thousand, (*ii*) the advertising budget to achieve a sales turnover of Rs. 200 lakh, and (*iii*) compute the coefficient of correlation.

(MBA, Delhi Univ., 2008)

27. The following data give the test scores and sales made by nine salesmen during the last one year :

Test Scores :	14	19	24	21	26	22	15	20	19
Sales ('000 Rs.) :	31	36	48	37	50	45	33	41	39

Obtain (*i*) the regression equation of test scores on sales, (*ii*) the regression equation of sales on test scores, and (*iii*) coefficient of correlation.
[$X = -2.312 + 0.5578\,Y$, (*ii*) $Y = 7.834 + 1.6083\,X$, (*iii*) $r = 0.947$]

28. A study of share prices of Textile group and Fertiliser group of companies yielded the following results :

	Textiles	*Fertilisers*
Mean	12.8	985.0
Standard Deviation	1.6	70.1
Coefficient of Correlation	+0.52	

The financial expert has estimated the likely price of textiles shares at the close of the next accounting year as 92. What would be your estimate of the likely price of fertiliser shares at the corresponding time ?

29. Following are the data on business turnover and staff of a company for eight years from 2003 to 2010 :

	2003	2004	2005	2006	2007	2008	2009	2010
Business Turnover (Rs. crore) :	45	50	60	75	80	110	150	170
Staff :	2,600	3,000	3,100	3,530	3,850	4,300	5,870	7,150

Fit a proper regression equation to estimate manpower in terms of business turnover. Estimate the staff requirement when the business turnover reaches Rs. 200 crore.

[$Y = 33.24X + 1100.3$; 7748.3]

30. The data on sales and promotion expenditure on a product for 10 years are given below :

Sales (Rs. lakh) :	8	10	9	12	10	11	12	13	14	15
Promotion Exp. (Rs. thousand) :	2	2	3	4	5	5	5	6	7	8

Use two-variable regression model to estimate the effect of promotion on sales. Forecast the sales for next year when the company hopes to spend Rs. 10 thousand on promotion.

[$X = 0.815\ Y - 4.591$, $Y = 1.003X + 6.686$, $Y_{10} = 16.716$]

31. Table below shows the power and top speeds of different brands of sports cars :

Brand :	*A*	*B*	*C*	*D*	*E*	*F*
Power X [kW] :	70	63	72	60	66	70
Speed Y [km/h] :	155	150	180	135	156	168
Brand :	*G*	*H*	*I*	*J*	*K*	*L*
Power X [kW] :	74	65	62	67	65	68
Speed Y [km/h] :	178	160	132	145	139	152

(*i*) Find the best linear relationship that fits the given data.

(*ii*) Estimate the speed of a car that has a power of 63 kW and find a 95% confidence interval for this estimate.

(*iii*) Determine how much of the variability in speed may be explained by the regression hypothesis.

32. 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 regression equations and find an estimate of Y which should correspond on an average to $X = 6.2$.

[$Y = 0.95X + 7.25$; $Y_{6.2} = 13.14$] *(MBA, Madurai-Kamaraj Univ., 2006)*

33. Family income and its percentage spent on food gave the following bivariate frequency table :

Food Expenditure	*Monthly Family Income (in hundred Rs.)*				
(in%)	*25–35*	*35–45*	*45–55*	*55–65*	*65–70*
15–20	8	9	12	13	8
20–25	6	3	6	11	14
25–30	—	7	9	—	4
30–35	5	8	10	14	13

(*i*) Estimate the family income for a food expenditure of 40%.

(*ii*) What amount should be spent on food expenditure for a monthly family income of Rs. 10,000.

(*iii*) Compute coefficient of correlation.

34. You are given below the following information about advertisement and sales.

	Adv. Exp. (X) *(Rs. crore)*	*Sales* (Y) *(Rs. crore)*
Mean	20	120
S.D.	5	25
Correlation coefficient	+0.8	

(*i*) Calculate the two regression equations.

(*ii*) Find the likely sales when advertisement expenditure is Rs. 25 crore.

(*iii*) What should be the advertisement budget if the company wants to attain sales target of Rs. 150 crore ?

[$Y = 4X + 40$; $X = 0.16Y + 0.8$; $Y_{25} = 140$; $X_{150} = 24.8$]

35. From the following data obtain the regression equation. Also find the correlation coefficient with the help of regression coefficient :

X:	6	2	10	4	8
Y:	9	11	5	8	7

[$Y = 11.9 - 0.65X$; $X = 16.4 - 1.3Y$, $r = -0.919$]

36. The monthly expenditure on advertisement and sales of a firm are given for 2010. It is generally found that expenditure on advertisement has its impact after two months. Allowing for time lag:

(*a*) calculate the correlation between expenditure on advertisement and sales.

(*b*) estimate the sales of the firm in February 2015.

Months/Year (2010)	*Expenditure on Advertisement (Rs.)*	*Sales (Rs.)*
January	50	1200
February	60	1500
March	70	1600
April	90	2000
May	120	2200
June	150	2500
July	140	2400
August	160	2600
September	170	2800
October	190	2900
November	200	3100
December	250	3900

37. The following figures relate to advertisement expenditure and sales :

Advertisement (in Rs. lakh) :	60	62	65	70	73	75	71
Sales (in Rs. crore) :	10	11	13	15	16	19	14

Estimate (*i*) the sales for advertisement expenditure of Rs. 80 lakh; and (*ii*) the advertisement expenditure for a sales target of Rs. 25 crore.

38. Given the regression equation of Y on X and X on Y are respectively $Y = 2X$ and $6X - Y = 4$ and the second moment of X about the origin is 3. Find (*i*) the correlation coefficient, and (*ii*) standard deviation of Y.

39. Find the regression coefficient of Y on X from the following regression equations :

$$5X = 22 + Y$$

$$64X = 24 + 45Y$$

Is it possible to calculate the standard deviation of Y from the given information ? Answer with reason.

40. A financial analyst has gathered the following data about the relationship between income and investment in securities in respect of 8 randomly selected families :

Income (Rs. '000) :	8	12	9	24	143	37	19	16
Per cent invested in securities :	36	25	33	15	28	19	20	22

(*a*) Develop an estimating equation that best describes these data.

(*b*) Find the coefficient of determination and interpret it.

(*c*) Calculate the standard error of estimate for this relationship.

(*d*) Find an approximate 90 per cent confidence interval for the percentage of income invested in securities by a family earning Rs. 25,000 annually.

41. From the data given below find :

(*i*) The two regression equations.

(*ii*) The coefficient of correlation between marks in Economics and Statistics.

(*iii*) The most likely marks in Statistics when the marks in Economics are 30.

Marks in Economics (X):	25	28	35	32	31	36	29	38	34	32
Marks in Statistics (Y) :	43	46	49	41	36	32	31	30	33	39

42. A financial analyst obtained the following information relating to return on security A and that of market portfolio M for the past 8 years :

Year :	1	2	3	4	5	6	7	8
Return on A :	10	15	18	14	16	16	18	4
Return on M :	12	14	13	10	9	13	14	7

(*i*) Develop an estimating equation that best describes these data.

(*ii*) Find the coefficient of determination and interpret it.

(*iii*) Determine the percentage of total variation in security return being explained by the return on the market portfolio.

(*MFC, Delhi Univ., 2005*)

43. Given the bivariate data :

X	:	1	5	3	2	1	1	7	3
Y	:	6	1	0	0	1	2	1	5

(*i*) Fit a regression equation of Y on X.

(*ii*) If a person has scored 8 on X variable, what would be his score on Y variable ?

44. Personnel Manager of a large industrial unit is interested to find a measure that can be used to fix the wages (yearly) of skilled workers. On experimental basis, the data on the length of service and their yearly wages (in Rs. '000) from a group of 10 randomly selected skilled workers are given below :

Length of service (X)	:	11	7	9	5	8	6	10	12	3	4
Yearly wages (Y)	:	14	11	10	9	13	10	14	16	6	7

(*a*) Develop the regression equation of wage (Y) on the length of service X.

(*b*) On the basis of (*a*) what initial pay the personnel manager should give to a skilled worker who has put in thirteen years of service on a similar basis, in another industry.

[$Y = 3.455 + 1.006\,X$; $Y = 16.533$] *(DIM, IGNOU, 2000)*

45. In a laboratory experiment on correlation research study, the equation to the two regression lines were to be $2X - Y + 1 = 0$ and $3X - 2Y + 7 = 0$. Find (*i*) the means of X and Y. Also work out the values of the regression coefficients and the coefficient of correlation between the two variables X and Y.

[$\overline{X} = 5$, $\overline{Y} = 11$; $bxy = 0.5$, $byx = 1.5$; $r = 0.866$]

46. An industrial engineer collected the following data on experience & performance rating of 8 operators :

Operators	:	1	2	3	4	5	6	7	8
Experience (years)	:	16	12	18	4	3	10	5	12
Performance Rating	:	87	88	89	68	58	80	70	85

(*a*) Does the data give evidence that experience improves performance ?

(*b*) Estimate the performance rating of an operator having (*a*) 9 years and (*b*) 15 years of experience.

[$Y = 69.67 + 1.133\,X$] *(MBA, Kumaun Univ., 2002)*

47. The following table gives the age of cars of certain make and the annual maintenance costs. Find (*i*) the coefficient of correlation between the variables and (*ii*) Regression equation for costs related to age.

Age of Cars (in years)	:	2	4	6	8	
Maintenance costs (in hundred Rs.)	:	0	20	25	30	*(MBA, HPU, 2002)*

48. A firm administers a test to sales trainees before they go into the field. The management of the firm is interested in determining the relationship between the test scores and the sales made by the trainees at the end of one year in the field. The following data were collected for ten sales personnel who have been in the field for one year :

Sales Person Number	*Test Score*	*Number of Units Sold*
1	2.6	95
2	3.7	140
3	2.4	85
4	4.5	180
5	2.6	100
6	5.0	195
7	2.8	115
8	3.0	136
9	4.0	175
10	3.4	150

(*i*) Find the regression line which would be used to predict sales from trainees test scores.

(*ii*) Predict the number of units which would be sold by trainee who received an average test score. *(MBA, DU, 2001)*

49. For the data given below :

	Average	*S.D.*
Production (in units)	35	10
Capacity utilisation (%)	85	8
Coefficient of correlation	0.6	

Obtain the two regression equations.

Estimate the production when the capacity utilisation is 70 per cent. *(MBA, D.U., 2003)*

50. Explain why there are two regression lines ? What happens if the two lines are identical ? For the data given below, find the relevant line of regression to estimate the price, if supply is 25 million tonnes.

Supply (m.t.)	:	5	10	12	15	18
Price (Rs./kg.)	:	16	15	12	12	10

(M. Com., A.M.U., 2001)

51. The following table shows the ages (X) and blood pressure (Y) of 8 persons :

X	:	52	63	45	36	72	65	47	25
Y	:	62	53	51	25	79	43	60	33

Obtain the regression equation of Y on X and find out the expected blood pressure of a person who is 49 years old.

(M.Com., Madurai-Kamaraj Univ., 2008)

52. Determine the equation of the straight line which best fits the following data :

X	Y
10	19
12	22
13	24
16	27
17	29
20	33
25	37

(MBA., IGNOU, 2005)

53. Regression calculations were carried out as follows :

$\sum X = 32, \sum Y = 24, \sum XY = 218$

$\sum X^2 = 296, \sum Y^2 = 162.5, n = 4$

Find the lines of regression and coefficient of correlation and comment. *(MBA, M.D. Univ., 2000)*

54. From the following data obtain the two regression equations :

Sales	:	91	97	103	121	67	124	52	73	111	57
Purchases	:	97	75	69	97	70	91	39	61	83	47

(MBA, Madurai Kamaraj Univ., Nov. 2001)

55. Obtain the regression of Y on X and X on Y from the following data and estimate the blood pressure when the age is 50.

Age	*Blood Pressure*	*Age*	*Blood Pressure*
50	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	150
47	128	60	155

(MBA, Bharathidasan Univ., 2001)

56. From the data given below, find the two regression equations and the most likely marks in statistics when marks in Economics are 30.

Marks in Economics	:	25	28	35	32	31	36	24	38	34	32
Marks in Statistics	:	43	46	49	41	36	32	31	30	33	39

(MBA, M.K. Univ., 2003)

57. Cost accountants often estimate overheads based on the level of production. At BFL company, the data collected are as follows. Find the best fit equation between production and overhead costs. Predict overheads when 50 units are produced.

Overhead	:	191	170	272	155	280	173	234	116	153	178
Production units	:	40	42	53	35	56	39	48	30	37	40

(MBA, Bharathidasan Univ., 2007)

*****

9

Index Numbers : Concepts and Applications

INTRODUCTION

Index numbers occupy a place of great prominence in business statistics. Though originally developed for measuring the effect of change in prices, there is hardly any field today where index numbers are not used. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies. In fact, they are described as *barometers of economic activity*, *i.e.*, if one wants to get an idea as to what is happening to an economy he should look to important indices like the index number of industrial production, agricultural production, business activity, etc.

Index numbers are playing an increasingly significant role in business planning and in the formulation of executive decisions. They are not directly used to prepare forecasts but many of the techniques employed in preparing forecasts utilize index numbers. For example, in correlation analysis either the dependent or independent variable or both may be in the form of index numbers. Many businesses are often reluctant to give out information concerning sales, profits, and the like. They may be induced to release some of the same data in the form of index numbers which permit the absolute value of this restricted information to be concealed. Under such conditions, it is possible to present index numbers indicating whether a firm's profits or sales have increased or decreased over a period of years without revealing the total amount of profits or sales.

An index number* may be described as a specialised average designed to measure the change in the level of a phenomenon with respect to time, geographic location or other characteristics such as income, etc. Thus, when we say that the index number of wholesale prices is 125 for the period Jan., 2010 compared to Jan., 2009 it means there is a net increase in the prices of wholesale commodities to the extent of 25 per cent.

For a proper understanding of the term index number, the following points are worth considering :

(1) *Index numbers are specialised averages*.As explained in the chapter on Measures of Central Tendency, an average is a single figure representing group of figures. However, to obtain an average the items must be comparable, for example, the average weight of men, women and children of a certain locality has no meaning at all. Furthermore, the unit of measurement must be the same for all the items. Thus an average of the weight expressed in kg., lb., etc., has no meaning. However, this is not so with index numbers. Index numbers are used for purposes of comparison in situations where two or more series are expressed in different units or the series are composed of different types of items. For example,

---

*An index number is a device which shows by its variation the changes in a magnitude which is not capable of accurate measurement in itself or of direct valuation in practice—Wheldom : *Business Statistics*.

"An index number represents the general level of magnitude of the changes between two or more situations of a number of variables taken as a whole."—*Karmel*

Index numbers are quantitative measures of the general level of growth of prices, production, inventory and other quantities of economic interest.—*Ronold*

while constructing a consumer price index the various items are divided into broad heads, namely (*i*) Food, (*ii*) Clothing, (*iii*) Fuel and Lighting, (*iv*) House Rent, and (*v*) Miscellaneous. These items are expressed in different units ; thus under the head 'food' wheat and rice may be quoted per quintal, ghee per kg., etc. Similarly, cloth may be measured in terms of metres. An average of all these items expressed in different units is obtained by using the technique of index numbers.

(2) *Index numbers measure the change in the level of a phenomenon.* Since index numbers are essentially averages they describe in one single figure the increase or decrease in the level of a phenomenon under study. If the index of industrial production is 115 in 2010 compared to 2009, it means that there is net increase in industrial production to the extent of 15%. It should be carefully noted that even where an index is showing a net increase, it may include some items which have actually decreased in value and others which have remained constant.

(3) *Index numbers measure the effect of changes over a period of time.* They are occasionally revised to take into account changes in the economy effected by technology, consumer tastes and spending patterns. Index numbers are most widely used for measuring changes over a period of time. Thus we can find out the net change in agricultural prices from the beginning of First Plan period to the end of the Eighth Plan period. Similarly, we can compare the agricultural production, industrial production, imports, exports, wages, etc., at two different times. However, it should be noted that index numbers not only measure change over a period of time but also compare economic conditions of different locations, different industries, different cities or different countries. But since the basic problems are essentially the same and since most of the important index numbers published by the Government and private research organisations refer to data collected at different times, we shall consider in this chapter index numbers measuring changes relative to time only. However, methods described can be applied to other cases also.

Uses of Index Numbers

Index numbers are indispensable tools of economic and business analysis. Their significance can be best appreciated by the following points :

(1) *They help in framing suitable policies.* Many of the economic and business policies are guided by index numbers. For example, for deciding the increase in dearness allowance of the employees, the employer has to depend primarily upon the cost of living index. If wages and salaries are not adjusted in accordance with the cost of living, very often it leads to strikes and lockouts which in turn cause considerable waste of resources.

Though index numbers are most widely used in the evaluation of business and economic conditions, there are a large number of other fields also where index numbers are useful. For example, sociologists may speak of population indices; psychologists measure intelligence quotients, which are essentially index numbers comparing a person's intelligence score with that of an average for his or her age; health authorities prepare indices to display changes in the adequacy of hospital facilities; and educational research organisations have devised formulae to measure changes in effectiveness of school systems.

(2) *They reveal trends and tendencies.* Since index numbers are most widely used for measuring changes over a period of time, the time series so formed enable us to study the general trend of the phenomenon under study. For example, by examining index numbers of imports for India for the last 8-10 years we can say that our imports are showing an upward tendency, *i.e.*, they are rising year after year. Similarly by examining the index numbers of industrial production, business activity, etc., for the last few years we can conclude about the trend of production and business activity. By examining the trends of the phenomenon under study we can draw very important conclusions as to how much change is taking place due to the effect of seasonality, cyclical forces, irregular forces, etc.

(3) *Index numbers are very useful in deflating.* Index numbers are used to adjust the original data for price changes, or to adjust wage for cost of living changes and thus transform nominal wages into real wages. Moreover, nominal income can be transformed into real income and nominal sales into real sales through appropriate index numbers. This point will be discussed in detail towards the end of the chapter.

Classification of Index Numbers

Index numbers may be classified in terms of what they measure. In economics and business the classifications are : (1) price; (2) quantity; (3) value; and (4) special purpose.*

Only price and quantity index numbers are discussed in detail. The others will be mentioned, but without details of how to construct them since both value and special purpose index numbers do not offer new problems in construction. Since the details of construction of all types of index numbers can be understood if the construction of price index numbers is understood, we shall devote major attention to them.

Problems in the Construction of Index Numbers

Before constructing index numbers a careful thought must be given to the following problems :

1. The purpose of the index. At the very outset the purpose of constructing the index must be very clearly decided—what the index is to measure and why? There is no all purpose index. Every index is of limited and particular use. Thus, a price index that is intended to measure consumers' prices must not include wholesale prices. And if such an index is intended to measure the cost of living of poor families, great care should be taken not to include goods ordinarily used by middle class and upper-income groups. Failure to decide clearly the purpose of the index would lead to confusion and wastage of time with no fruitful results. All other problems such as the base year, the number of commodities to be included, the prices of the commodities, etc., are decided in the light of the purpose for which the index is being constructed.

The problem of the scope of the index, *i.e.,* the field covered by the index, is linked up with the purpose of the index and the data available. The data available, or rather the lack of them, may necessitate the modification of the purpose.

2. Availability and comparability of data. It is needless to say that it is impossible to make appropriate comparisons unless the necessary statistical data can be obtained. Many persons while constructing an index have been frustrated by the fact that essential information was tabulated by countries, whereas actually they needed it by townships, they have run into difficulties because sales data were available only by type of merchandise and not by brand.

The problem of comparability of data used in an index can also be quite troublesome. It is an exceedingly difficult problem to make sure that prices are actually comparable that they really refer to goods and services that are identical in quality. The comparability of statistical data may also be questioned if parts of the data were collected by different agencies. Mistakes in the selection of data that are really not comparable, can also be made at times due to the carelessness of the persons constructing the index.

To summarise it is important to keep in mind that, in so far as is possible, data which are used in the construction of an index number must be comparable in the sense that if one wants to compare prices one is not really of comparing quality. Furthermore, the goods or services to which the prices or quantities refer must adhere to uniform definitions that is, rigorous specifications. How to achieve these goals in practice is a problem that has never been solved entirely to everyone's satisfaction.

---

* Index numbers may be constructed for a single commodity, called *simple index numbers,* or for a group of commodities called *composite index numbers.*

3. Selection of base period. Whenever index numbers are constructed, a reference is made to some base period. The base period is the period with which comparisons of relative changes are made. It may be a year, month or a day. The index for base period is always taken as 100. Though the selection of the base period would primarily depend upon the object of the index, the following points need careful consideration in the selection of base period:

(*i*) *The base period should be normal one.* The period that is selected as base should be normal, *i.e.*, it should be free from abnormalities like wars, earthquakes, famines, booms, depressions, etc. However, at times it is really difficult to select a year which is normal in all respects—a year which is normal in one respect may be abnormal in another. To solve this problem an average of a number of years, say, 3 or 4 (preferably covering one complete cycle), may be taken as the base. The process of averaging will reduce the effect of extremes. Thus the average of the period from 2006 to 2010 may be considered normal whereas no individual year in that span may be considered normal.

(*ii*) *The base period should not be too distant in the past.* Since index numbers are helpful in decision-making and economic policies are often a matter of short period, we should not select a base period that is too distant in the past. For example, for deciding increase in dearness allowance at present, there is no advantage in taking 1995 or 2000 as the base ; the comparison should be with the preceding year or the year after which dearness allowance has not been revised.

(*iii*) *Fixed base or chain base.* While selecting the base decision has to be made as to whether the base shall remain fixed or not, *i.e.*, whether we have a fixed base or chain base index. In the fixed base method, the year or the period of years to which all other prices are related is constant for all times. On the other hand, in the chain base method the prices of a year are linked with those of the preceding year and not with the fixed year. Naturally the chain base method gives a better picture than what is obtained by fixed base method. However, much would depend upon the purpose of constructing the index.

4. Selection of number of items. Every item cannot be included while constructing an index number and hence one has to select a sample. For example, while constructing a price index it is impossible to include each and every commodity. Hence, it is necessary to decide what commodities to include. The commodities should be selected in such a manner that they are representative of the tastes, habits and customs of the people for whom the index is meant. Thus in a consumer price index for working class, items like colour T.V., motor cars, refrigerators, cosmetics, etc., find no place. A decision must also be made on the number of commodities to be included and their qualities. Here we should note that the larger the number of commodities included, the more representative shall be the index but at the same time the greater shall be the cost and the time taken. The purpose of the index shall help in deciding the number of commodities. Thus, in a general price index a larger number of commodities shall have to be included as compared to a specific purpose index such as the index number of the prices of foodgrains or industrial raw materials.

It is also necessary to decide the grade or quality of the items to be included in the index. Index numbers shall give wrong result if at one time one set of qualities is included and at another time another set. To avoid confusion about qualities it is desirable that as far as possible no standardised or graded items are included so that they can be easily identified after a time lapse.

5. Price quotations. After the commodities have been selected, the next problem is to obtain price quotations for these commodities. It is a well-known fact that prices of many commodities vary from place to place and even from shop to shop in the same market. It is impracticable to obtain price quotations from all the places where a commodity is dealt with. A selection must be made of representative places and persons. These places should be those which are well-known for trading for that particular commodity. After the places from where the price quotations are to be obtained are decided, the next thing is to appoint some person or institution who can supply price quotations as and when required. Great care must be exercised to see that the price reporting agency is unbiased. In order to check the accuracy of price quo-

tations supplied by an agency is that quotations are obtained from more than one agency. If there is some reliable journal or magazine supplying price quotations then it may be utilised.

In order to ensure uniformity the manner in which prices are to be quoted must also be decided. There are two methods of quoting prices : (*i*) money prices, and (*ii*) quantity prices. In the former case prices are quoted per unit of commodity, for example, sugar Rs. 2600 per quintal (100 kg.) and in the latter case prices are quoted per unit of money. Thus, sugar may be quoted as 1/2 kg. for thirteen rupees. The former method is free from confusion and is generally adopted while quoting prices.

A decision must also be made as to whether the wholesale prices or retail prices are required. The choice would depend upon the purpose of the index. Thus in a consumer price index, the wholesale price shall not be representative at all. If the prices of certain commodities are controlled by the government then it is these controlled prices that should be taken into account and not the black-market prices, which may be much higher.

6. Choice of an average. Since index numbers are specialized averages, a decision has to be made as to which particular average (*i.e.*, arithmetic mean, median, mode, geometric mean or harmonic mean) should be used for constructing the index. Median, mode and mean are almost never used in constructing the index numbers. Basically a choice has to be made between arithmetic mean and geometric mean. Theoretically speaking, geometric mean is the best average in the construction of index numbers because of the following reasons : (*i*) in the construction of index numbers we are concerned with ratios or relative changes and the geometric mean gives equal weights to equal ratio of change; (*ii*) geometric mean is less susceptible to major variations as a result of violent fluctuations in the values of the individual items; and (*iii*) index numbers calculated by using this average are reversible and, therefore, base shifting is easily possible. The geometric mean index always satisfies the time reversal test.

Despite theoretical justification for favouring geometric mean, arithmetic mean is more popularly used while constructing index numbers. This is for the reason that arithmetic mean is much more simple to compute than the geometric mean. However, wherever possible, in the interest of greater accuracy geometric mean should be preferred. It is gratifying to note that with the growing use of calculating devices, geometric mean is becoming more popular in constructing index numbers.

7. Selection of appropriate weights. The problem of selecting suitable weights is quite important and at the same time quite difficult to decide. The term 'weight' refers to the relative importance of the different items in the construction of the index. All items are not of equal importance and hence it is necessary to devise some suitable method whereby the varying importance of the different items is taken into account. This is done by allocating weights. Thus we have broadly two types of indices—unweighted indices and weighted indices. In the former case, no specific weights are assigned, whereas in the latter case specific weights are assigned to various items. It may be pointed out here that no index is unweighted in strict sense of the term as weights implicitly enter in unweighted indices because we are giving equal importance to all the items and hence weights are unity. It is, therefore, necessary to adopt some suitable method of weighting so that arbitrary and haphazard weights may not affect the results.

There are two methods of assigning weights: (*i*) implicit, and (*ii*) explicit.

In the implicit weighting, a commodity or its variety is included in the index a number of times. Thus, if wheat is to be given in an index twice as much weight as rice, then two varieties of wheat as against one rice may be included in the series. On the other hand, in case of implicit weighting some outward evidence of importance of the various items in the index is given. When the explicit weights are assigned the questions are: (*i*) By what do we weight? and (*ii*) What types of weight do we use?

(*i*) In order to bring out the economic importance of the commodities involved the weight can be production figures, consumption figures or distribution figures.

(*ii*) Weights are of two types: quantity weights and value weights. A quantity weight, symbolised by q, means the amount of commodity produced, distributed, or consumed in some time period. A value weight, on the other hand, combines price with quantity 'produced, distributed or consumed'. Value is in terms of rupees and is symbolised by $p \times q$ where p stands for the price and q for the quantity.

Now the question is whether to choose quantity weights or value weights. The statistician is not free to choose here. If the aggregative method is used while constructing index, then quantities are used as weights because price times quantity will always give the same units, namely, rupees. On the other hand, in averaging price relatives quantity figures cannot be used. It is for the reason that if we multiply percentages by quantities expressed in different units, we get result in different units; for example, percentage tonnes will give tonnes and percentages multiplied by kg. will give kg. Such figures cannot be used in computation. But if percentages are multiplied by value figures which are always expressed in rupees, we get answer in rupees only. Hence the statistician will use q as a weight in the method of aggregating actual prices and must use $p \times q$ as a weight in the method of averaging price relatives.*

Another problem in connection with weights is that of deciding whether the weights shall be fixed or fluctuating. Since the relative importance of the different items does not remain the same for all times, it is logical to vary the weights from time to time. Such an index would give better results. However, when fluctuating weights are used one must be very careful in interpreting the index because not only changes in prices but also changes in weights are affecting the index.

One of the outstanding problems in index number construction is that of devising a weighting system that will accurately represent the commodities throughout the period covered by the index number. Many systems have been tried, such as getting the average importance of the commodities over a period of years, but no perfect system has yet been developed.

8. Selection of an appropriate formula. A large number of formulae have been devised for constructing the index. The problem very often is that of selecting the most appropriate formula. The choice of the formula would depend not only on the purpose of the index but also on the data available. Prof. Irving Fisher has suggested that an appropriate index is that which satisfies time reversal test and factor reversal test. Theoretically, Fisher's method is considered as 'ideal' for constructing index numbers. However, from a practical point of view there are certain limitations of this index which shall be discussed later. As such, no particular formula can be regarded as the best under all circumstances. On the basis of his knowledge of the characteristics of different formulae a discriminating investigator will choose technical methods adapted to his data and appropriate to his purposes.

None of the above problems is simple to solve in practice and the final index is usually the product of compromise between theoretical standards and the standards attainable with the given data.

METHODS OF CONSTRUCTING INDEX NUMBERS

A large number of formulae have been devised for constructing index numbers. Broadly speaking, they can be grouped under two heads:

(*a*) Unweighted indices; and

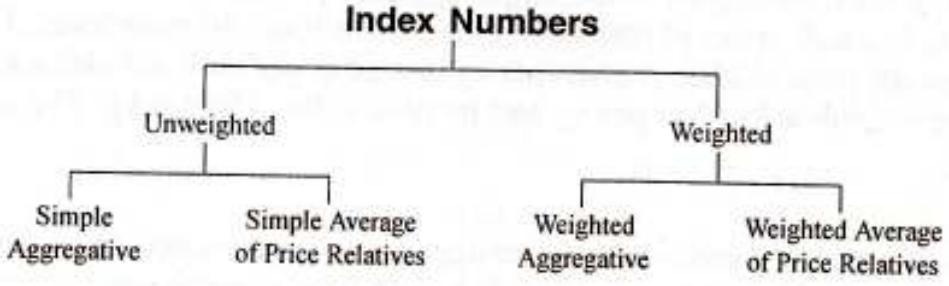
(*b*) Weighted indices.

In the unweighted indices, weights are not expressly assigned whereas in the weighted indices, weights are assigned to the various items. Each of these types may further be divided under two heads:

(*i*) Simple Aggregative; and

(*ii*) Simple Average of Price Relatives.

The following chart illustrates the various methods :



*Sometimes in the absence of actual weights arbitrary magnitudes may have to be used as weights. However, it is unscientific to use these weights and, therefore, they should be only in the crudest forms of index numbers.

A. UNWEIGHTED INDEX NUMBERS

I. Simple Aggregative Method

This is the simplest method of constructing index numbers. When this method is used to construct a price index, the total of current year prices for the various commodities in question is divided by the total of base year prices and the quotient is multiplied by 100. Symbolically,

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

where Σp_1 = Total of current year prices for various commodities, and
Σp_0 = Total of base year prices for various commodities.

This method of constructing the index is very simple and the steps required in computation are :

(*i*) Add the current year prices for various commodities, *i.e.*, obtain Σp_1.

(*ii*) Add the base year prices for the same commodities, *i.e.*, obtain Σp_0.

(*iii*) Divide Σp_1 by Σp_0 and multiply the quotient by 100.

Illustration 1. From the following data construct an index number for 2010 taking 2009 as base :

Commodity and unit	*Price (Rs.) 2009*	*Price (Rs.) 2010*
Butter (kg.)	*110.00*	*120.00*
Cheese (kg.)	75.00	80.00
Milk (lt.)	13.00	13.00
Bread (1)	9.00	9.00
Eggs (Doz.)	18.00	20.00
Ghee (1 tin)	850.00	860.00

Solution. CONSTRUCTION OF PRICE INDEX

Commodity	*Price in 2009* p_0	*Price in 2010* p_1
Butter (kg.)	110.00	120.00
Cheese (kg.)	75.00	80.00
Milk (lt.)	13.00	13.00
Bread (1)	9.00	9.00
Eggs (Doz.)	18.00	20.00
Ghee (1 tin)	850.00	860.00
	$\Sigma p_0 = 1075.00$	$\Sigma p_1 = 1102.00$

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{1102}{1075} \times 100 = 102.51$$

This means that as compared to 2009, in 2010 there is a net increase in price of commodities included in the index to the extent of 2.51%.

Limitations of the Method

There are two main limitations of the simple aggregative index :

(1) The units in which prices of commodities are given affect the price index. For example, if in the above illustration the price of ghee is given per kg instead of per tin it will make a difference. Suppose the price of ghee in 2009 is Rs. 200 per kg and in 2010 is Rs. 250 per kg. The index would then be $\frac{250}{200} \times 100 = 125$.

(2) No consideration is given to the relative importance of the commodities. The unit by which each item happens to be priced introduces an implicit weight. This concealment is undesirable and severely restricts the usefulness of an index number arrived at through the method of simple aggregate of actual prices.

II. Simple Average of Relatives Method

When this method is used to construct a price index, price relatives are obtained for the various items included in the index and then an average of these relatives is obtained using any one of the measures of central tendency, *i.e.*, arithmetic mean, median, mode, geometric mean or harmonic mean. When arithmetic mean is used for averaging the relatives, the formula for computing the index is :

$$P_{01} = \frac{\Sigma\left(\frac{p_1}{p_0} \times 100\right)}{N}$$

where N refers to the number of items (commodities) whose price relatives are thus averaged.

Although any measure of central tendency can be used to obtain the overall index, price relatives are generally averaged either by the arithmetic or the geometric mean. When geometric mean is used for averaging the price relatives, the formula for obtaining the index becomes

$$\log P_{01} = \frac{\Sigma \log\left(\frac{p_1}{p_0} \times 100\right)}{N} \text{ or } \frac{\Sigma \log P}{N}$$

where $$P = \frac{p_1}{p_0} \times 100$$

or, $$P_{01} = \text{antilog}\left[\frac{\left(\Sigma \log \frac{p_1}{p_0}\right) \times 100}{N}\right] = \text{antilog}\left(\frac{\Sigma \log P}{N}\right)$$

Other measures of central tendency are not in common use for averaging relatives.

Illustration 2. From the data of Illustration 1, compute price index by simple average of price relatives method based on (*a*) arithmetic mean, and (*b*) geometric mean.

Solution. (*a*) PRICE INDEX BASED ON SIMPLE AVERAGE OF PRICE RELATIVES

Commodities	*Price (Rs.) 2009* (p_0)	*Price (Rs.) 2010* (p_1)	$\frac{p_1}{p_0} \times 100$
Butter (kg.)	110.00	120.00	109.09
Cheese (kg.)	75.00	80.00	106.67
Milk (lt.)	13.00	13.00	100.00
Bread (1)	9.00	9.00	100.00
Eggs (Doz.)	18.00	20.00	111.11
Ghee (1 tin)	850.00	860.00	101.18
$N = 6$			$\Sigma \frac{p_1}{p_0} \times 100 = 628.05$

$$\text{Price Index or } P_{01} = \frac{\Sigma \frac{p_1}{p_0} \times 100}{N} = \frac{628.05}{6} = 104.67.$$

(*b*) PRICE INDEX BASED ON GEOMETRIC MEAN OF PRICE RELATIVES

Commodities	*Price (Rs.) 2009* (P_0)	*Price (Rs.) 2010* (P_1)	*Price Relatives P*	*log P*
Butter (kg.)	110.00	120.00	109.09	2.0378
Cheese (kg.)	75.00	80.00	106.67	2.0280
Milk (lt.)	13.00	13.00	100.00	2.0000
Bread (1)	9.00	9.00	100.00	2.0000
Eggs (doz.)	18.00	20.00	111.11	2.0457
Ghee (1 tin)	850.00	860.00	101.18	2.0051
				$\Sigma \log P = 12.1166$

$$P_{01} = AL\left[\frac{\Sigma \log P}{N}\right] = AL\left[\frac{12.1166}{6}\right] = AL\ 2.0194 = 104.57$$

Although arithmetic mean and geometric mean have both been used, the arithmetic mean is often preferred because it is easier to compute and much better known. Some economists, notably F.Y. Edgeworth, have preferred to use the median which is not affected by single extreme value. Since the argument is important only when an index is based on a very small number of commodities, it generally does not carry much weight and the median is seldom used in actual practice.

Merits and Limitations of this Method

Merits. This method has the following two advantages over the previous method:

1. Extreme items do not influence the index. Equal importance is given to all the items.

2. The index is not influenced by the units in which prices are quoted or by the absolute level of individual prices. Relatives are pure numbers and are, therefore, independent of the original units. Consequently, index numbers computed by the relative method would be the same regardless of the way in which prices are quoted.

Limitations. Despite these merits this method is not very satisfactory because of the following two reasons :

1. Difficulty is faced with regard to the selection of an appropriate average. The use of the arithmetic mean is considered as questionable sometimes because it has an upward bias. The use of geometric mean involves difficulties of computations. Other averages are almost never used while constructing index numbers.

2. The relatives are assumed to have equal importance. This is again a kind of concealed weighting system that is highly objectionable since economically some relatives are more important than others.

B. WEIGHTED INDEX NUMBERS

The unweighted index numbers discussed so far are not unweighted in the true sense of the term. They assign equal importance to all the items included in the index and as such they are in reality weighted, weights being implicit rather than explicit. As discussed earlier, in case of unweighted indices it is possible to get different results by changing the importance of different items by quoting prices relative to different units. Implicit weighting (or the unweighted index) is far from realistic in most of the cases. Construction of useful index numbers requires a conscious effort to assign to each commodity a weight in accordance with its importance in the total phenomenon that the index is supposed to describe.

Weighted index numbers are of two types:

I. Weighted Aggregative Index Numbers, and

II. Weighted Average of Relative Index Numbers.

I. Weighted Aggregative Index Numbers

These index numbers are of the simple aggregative type with the fundamental difference that weights are assigned to the various items included in the index. There are various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of which some of the more important ones are :

1. Laspeyres method,
2. Paasche method,
3. Dorbish and Bowley's method,
4. Fisher's Ideal method,
5. Marshall-Edgeworth method, and
6. Kelly's method.

All these methods carry the name of persons who have suggested them.

1. Laspeyres Method. In this method the base year quantities are taken as weights. The formula for constructing index is:

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$$

Steps. (*i*) Multiply the base year prices of various commodities with base year weights and obtain $\Sigma p_1 q_0$.

(*ii*) Multiply the base year prices of various commodities with base year weights and obtain $\Sigma p_0 q_0$.

(*iii*) Divide $\Sigma p_1 q_0$ by $\Sigma p_0 q_0$ and multiply the quotient by 100. This gives us the price index.

Laspeyres index attempts to answer the question: "What is the change in aggregate value of the base period list of goods when valued at given period prices?" The index is very widely used in practical work.

2. Paasche Method. In this method the *current year* quantities are taken as weights. The formula for constructing the index is:

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Steps. (*i*) Multiply the current year prices of various commodities with current year weights and obtain $\sum p_1 q_1$.

(*ii*) Multiply the base year prices of various commodities with current year weights and obtain $\sum p_0 q_1$.

(*iii*) Divide $\Sigma p_1 q_1$ by $\Sigma p_0 q_1$ and multiply the quotient by 100.

In general this formula answers the question: "What would be the value of the given period list of goods when valued at base period prices?"

Comparison of Laspeyres and Paasche methods. From a practical point of view, Laspeyres index is often preferred to Paasche's for the simple reason that in Laspeyres index weights (q_0) are the base year quantities and do not change from one year to the next. On the other hand, the use of Paasche index requires the continuous use of new quantity weights for each period considered and in most cases these weights are difficult and expensive to obtain.

An interesting property of Laspeyres and Paasche indices is that the former is generally expected to *overestimate,* or to leave an upward bias, whereas the latter tends to *underestimate, i.e.,* show a downward bias. When the prices increase, there is usually a reduction in the consumption of those items for which the increase has been the most pronounced, and, hence, by using base year quantities we will be giving too much weight to the prices that have increased the most and the numerator of the Laspeyres index will be too large. When the prices go down, consumers often shift their preference to those items which have declined the most and hence, by using base period weights in the numerator of the Laspeyres index we shall not be giving sufficient weight to the prices that have gone down the most and the numerator will again be too large. Similarly because people tend to spend less on goods when their prices are rising the use of the Paasche or current weighting produces an index which tends to underestimate the rise in prices, *i.e.,* it has a downward bias. But the above arguments do not imply that Laspeyres index must necessarily be larger than the Paasche index.

Unless drastic changes have taken place between the base year and the given year, the difference between the Laspeyres and Paasche's will generally be small and either could serve as a satisfactory measure. In practice, however, the base year weighted Laspeyres type index remains the most popular for reasons of its practicability. The Paasche type index can only be constructed when up-to-date data for the weights are available. Furthermore, the price index of a given year can be compared only with the

base year. For example, let $P_{2006} = 100, P_{2007} = 130$, and $P_{2008} = 140$. Then P_{2007} and P_{2008} are using different weights and cannot be compared with each other. If these indices had been obtained by the Laspeyres formula, they could be compared because in that case the weights are the same base year weights (q_0). For these reasons, in practice the Paasche formula is usually not used and the Laspeyres type index remains most popular for reasons of its practicability.

3. Dorbish and Bowley's Method. Dorbish and Bowley have suggested simple arithmetic mean of the two indices (Laspeyres and Paasche) mentioned above so as to take into account the influence of both the periods, *i.e.,* current as well as base periods. The formula for constructing the index is:

$$P_{01} = \frac{L + P}{2}$$

where L = Laspeyres Index, P = Paasche Index

or

$$P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$$

4. Fisher's 'Ideal' Method. Prof. Irving Fisher has given a number of formulae for constructing index number and of these he calls one as the 'ideal' index. The Fisher's Ideal Index is given by the formula* :

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

or

$$P_{01} = \sqrt{L \times P}$$

It shall be clear from the above formula that Fisher's Ideal Index is the geometric mean of the Laspeyres and Paasche indices.

The above formula is known as 'Ideal' because of the following reasons:

(*i*) It is based on the geometric mean which is theoretically considered to be the best average for constructing index numbers.

(*ii*) It takes into account both current year as well as base year prices and quantities.

(*iii*) It satisfies both the time reversal test as well as the factor reversal test as suggested by Fisher.

(*iv*) It is free from bias. The two formulae (Laspeyres' and Paasche's) that embody the opposing types and weight biases are, in the ideal formula, crossed geometrically, *i.e.*, by an averaging process that of itself has no bias. The result is the complete cancellation of biases of the kinds revealed by time reversal and factor reversal tests.

It is not, however, a practical index to compute because it is excessively laborious. The data, particularly or the Paasche segment of the index, are not readily available. In practice, statisticians will continue to rely upon simple, although perhaps less exact, index number formulae.

5. Marshall-Edgeworth Method. In this method also both the current year as well as base year prices and quantities are considered. The formula for constructing the index is :

$$P_{01} = \frac{\Sigma (q_0 + q_1) p_1}{\Sigma (q_0 + q_1) p_0} \times 100$$

or opening the brackets

*For proof see under Tests For Perfection.

$$P_{01} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

It is a simple, readily constructed measure, giving a very close approximation to the results obtained by the ideal formula.

6. Kelly's Method. T.L. Kelly has suggested the following formula for constructing index number :

$$P_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

Here weights are the quantities which may refer to some period, not necessarily the base year or current year. Thus, the average quantity of two or more years may be used as weights. If in the Kelly's formula the average of the quantities of two years is used as weights, the formula becomes

$$P_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

where

$$q = \frac{q_0 + q_1}{2}$$

Similarly, the average of the quantities of three or more years can be used as weights. The method is known as *fixed weight aggregative index* and is currently in great favour in the construction of index number series. An important advantage of this formula is that like Laspeyres' index it does not demand yearly changes in the weights. Moreover, the base period can be changed without necessitating corresponding change in the weights. This is very important because the construction of appropriate quantity weights for general purpose index usually requires a considerable amount of work. Weights can thus be kept constant until new census (or other survey) data become available to revise the index.

Illustration 3. Construct index numbers of price from the following data by applying :

1. Laspeyres' method,
2. Paasche's method,
3. Bowley's method,
4. Fisher's method, and
5. Marshall-Edgeworth method.

Commodity	*2009*		*2010*	
	Price	*Quantity*	*Price*	*Quantity*
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution.

CALCULATION OF VARIOUS INDICES

Commodity	*2009*		*2010*					
	Price p_0	*Quantity* q_0	*Price* p_1	*Quantity* q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	2	8	4	6	32	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	70	56	50	40
D	2	19	2	13	38	38	26	26
					$\Sigma p_1q_0 = 200$	$\Sigma p_0q_0 = 160$	$\Sigma p_1q_1 = 130$	$\Sigma p_0q_1 = 103$

1. *Laspeyres Method :*

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{200}{160} \times 100 = 125.$$

2. *Paasche's Method :*

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{130}{103} \times 100 = 126.21.$$

3. *Bowley's Method :*

$$P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100 = \frac{\frac{200}{160} + \frac{130}{103}}{2} \times 100$$

$$= \frac{1.25 + 1.2621}{2} \times 100 = \frac{2.5121}{2} \times 100 = 125.605$$

or

$$P_{01} = \frac{L + P}{2} = \frac{125 + 126.21}{2} = 125.605$$

4. *Fisher's Ideal Method :*

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100$$

$$= \sqrt{1.578} \times 100 = 1.256 \times 100 = 125.6.$$

5. *Marshall-Edgeworth Method :*

$$P_{01} = \frac{\Sigma (q_0 + q_1) p_1}{\Sigma (q_0 + q_1) p_0} \times 100 = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

$$= \frac{200 + 130}{160 + 103} \times 100 = \frac{330}{263} \times 100 = 125.475$$

II. Weighted Average of Relative Index Numbers

In the weighted aggregative methods discussed above price relatives were not computed. However, like unweighted relatives method it is also possible to compute weighted average of relatives. For purposes of averaging we may use either the arithmetic mean or the geometric mean. The **steps** in the computation of the weighted arithmetic mean of relatives index number are as follows :

(*i*) Express each item of the period for which the index number is being calculated as a percentage of the same item in the base period.

(*ii*) Multiply the percentages as obtained in step (*i*) for each item by the weight which has been assigned to that item.

(*iii*) Add the results obtained from the several multiplications carried out in step (*ii*).

(*iv*) Divide the sum obtained in step (*iii*) by the sum of the weights used. The result is the index number. Symbolically :

$$P_{01} = \frac{\Sigma PV}{\Sigma V}$$

where P = Price relative

V = Value weights, *i.e.*, $p_0 q_0$.

Instead of using arithmetic mean the geometric mean may be used for averaging relatives. The weighted geometric mean of relatives is computed in the same manner as the unweighted geometric mean of relatives index number except that weights are introduced by applying them to the logarithms of the relatives. When this method is used the formula for computing the index is :

$$P_{01} = \frac{\Sigma V . \log P}{\Sigma V}$$

where $P = \frac{p_1}{p_0} \times 100$

and V^* = Value weight, *i.e.*, $p_0 q_0$ for each item.

*If current year values are employed, the weights are $p_1 q_1$. If theoretical values are used as weights, the weights are $p_1 q_0$ or $p_0 q_1$.

Steps : (*i*) Obtain percentage relatives for each item.

(*ii*) Find the logarithm of each percentage relative found in step (*i*).

(*iii*) Multiply the logarithms by weights assigned.

(*iv*) Add the results obtained in step (*iii*).

(*v*) Divide the total obtained in step (*iv*) by the sum of the weights.

(*vi*) Find the antilogarithm of the quotient obtained in step (*v*). This is weighted geometric mean of relatives index number.

Illustration 4. From the following data compute price index by applying weighted average of price relatives method using : (*a*) arithmetic mean, and (*b*) geometric mean.

Commodity	p_0 (*Rs.*)	q_0	p_1 (*Rs.*)
Sugar	18.00	20 kg.	20.00
Flour	12.00	40 kg.	14.00
Milk	15.00	10 lt.	16.00

Solution. (*a*) INDEX NUMBER USING WEIGHTED ARITHMETIC MEAN OF PRICE RELATIVES

Commodity	p_0 (*Rs.*)	q_0	p_1 (*Rs.*)	p_0q_0 V	$\frac{p_1}{p_0} \times 100$ P	PV
Sugar	18.00	20 kg.	20.00	360	111.11	39999.6
Flour	12.00	40 kg.	14.00	480	116.67	56001.6
Milk	15.00	10 lt.	16.00	150	106.67	16000.5
				$\Sigma V = 990$		$\Sigma PV = 112001.7$

$$P_{01} = \frac{\Sigma PV}{\Sigma V} = \frac{112001.7}{990} = 113.13$$

This means that there has been a 13.13 per cent increase in price over the base level.

(*b*) INDEX NUMBER USING GEOMETRIC MEANS OF PRICE RELATIVES

Commodity	p_0 (*Rs.*)	q_0	p_1 (*Rs.*)	V	P	*Log P*	*V. Log P*
Sugar	18.00	20 kg.	20.00	360	111.11	2.046	736.56
Flour	12.00	40 kg.	14.00	480	116.67	2.067	992.16
Milk	15.00	10 lt.	16.00	150	106.67	2.028	304.20
				$\Sigma V = 990$			$\Sigma V.\ log\ P = 2032.92$

$$P_{01} = \text{A.L.}\left[\frac{\Sigma V.\log P}{\Sigma V}\right] = \text{A.L.}\left[\frac{2032.92}{990}\right] = \text{A.L. } 2.0535 = 113.11$$

The result obtained by applying the Laspeyres method would come out to be the same as obtained by weighted arithmetic mean of price relatives method (as shown below):

PRICE INDEX BY LASPEYRES' METHOD

Commodity	p_0 (*Rs.*)	q_0	p_1 (*Rs.*)	p_1q_0	p_0q_0
Sugar	18.00	20 kg.	20.00	400	360
Flour	12.00	40 kg.	14.00	560	480
Milk	15.00	10 lt.	16.00	160	150
				$\Sigma p_1q_0 = 1120$	$\Sigma p_0q_0 = 990$

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{1120}{990} \times 100 = 113.13$$

The answer is almost the same as obtained by weighted arithmetic mean of price relatives method.

Merits of Weighted Average of Price Relatives Method

The following are the special advantages of weighted average of relative indices over weighted aggregative indices :

(1) When different index numbers are constructed by the average of price relatives method, all of which have the same base, they can be combined to form a new index.

(2) When an index is computed by selecting one item from each of the many sub-groups of items, the values of each sub-group may be used as weights. Then only the method of weighted average of relatives is appropriate.

(3) When a new commodity is introduced to replace the one formerly used, the relative for the new item may be spliced to the relative for the old one, using the former value weights.

(4) The price or quantity relatives for each single item in the aggregate are, in effect, themselves a simple index that often yields valuable information for analysis.

Quantity Index Numbers

Price index numbers measure and permit comparison of the price of certain goods, quantity index numbers. On the other hand, measure and permit comparison of the physical volume of goods produced or distributed or consumed. Though price indices are more widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.

In constructing quantity index numbers, the problems confronting the statistician are analogous to those involved in price indices. We measure changes in quantities, and when we weigh we use prices or values as weights. Quantity indices can be obtained easily by changing p to q and q to p in the various formulae discussed above.

Thus, when Laspeyres' method is used

$$Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$$

When Paasche's formula is used

$$Q_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100$$

When Fisher's formula is used

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$$

These formulae represent the *quantity index* in which the quantities of the different commodities are weighted by their prices. However, any other suitable weights can be used instead.

Illustration 5. Compute by suitable method the index number of quantity from the data given below:

	2009		*2010*	
Commodity	*Price*	*Value*	*Price*	*Value*
A	8	80	10	110
B	10	90	12	108
C	16	256	20	340

Solution. Since we are given the value and the price we can obtain quantity figure by dividing value by price for each commodity. We can then apply Fisher's method for finding out quantity index.

Commodity	2009		2010		q_1p_0	q_0p_0	q_1p_1	q_0p_1
	p_0	q_0	p_1	q_1				
A	8	10	10	11	88	80	110	100
B	10	9	12	9	90	90	108	108
C	16	16	20	17	272	256	340	320
					$\Sigma q_1p_0 = 450$	$\Sigma q_0p_0 = 426$	$\Sigma q_1p_1 = 558$	$\Sigma q_0p_1 = 528$

$$\text{Quantity index or } Q_{01} = \sqrt{\frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times \frac{\Sigma q_1p_1}{\Sigma q_0p_1}} \times 100 = \sqrt{\frac{450}{426} \times \frac{558}{528}} \times 100$$

$$= \sqrt{1.116} \times 100 = 1.056 \times 100 = 105.6$$

Volume Index Numbers

The value of single commodity is the product of its price and quantity. Thus a value index V is the sum of the values of a given year divided by the sum of the values of the base year. The formula, therefore, is

$$V = \frac{\Sigma p_1q_1}{\Sigma p_0q_0} \times 100$$

where Σp_1q_1 = Total value of all commodities in the given period and Σp_0q_0 = Total value of all commodities in the base period.

Since in most cases the value figures are given, the formula can be stated more simply

$$V = \frac{\Sigma V_1}{\Sigma V_0}$$

in which V stands for value.

In this type of index both price and quantity are variable in the numerator. Weights do not have to be applied, since they are inherent in the value figures. A value index, therefore, is an aggregate of values. It measures the change in actual values between the base and the given periods.

The value index is not in wide use, although because of the unsatisfactory nature of price and quantity indices, it has been occasionally suggested that they be replaced by the value index. The temptation, however, must be resisted, since the concepts of price level and quantity level answer questions that cannot be answered by the value level. Furthermore, an aggregate of values may be viewed as the product of a price level and quantity level. The division of an aggregate of value into its price and quantity factors may be arbitrary, but this need not create any confusion of thought as long as our concepts of the two factors are consistent.

The test of consistency is that the product of the price and quantity indices must produce the value index.

TESTS FOR PERFECTION

Several formulae have been suggested for constructing index numbers and the problem is that of selecting the most appropriate one in a given situation. The following tests are suggested for choosing an appropriate index:

1. Time Reversal Test, 2. Factor Reversal Test, and
3. Circular Test.

1. Time Reversal Test

Prof. Irving Fisher had made a careful study of the various proposals for computing index numbers and has suggested various tests to be applied to any formula to indicate whether or not it is satisfactory. The two most important of these he calls the time reversal test and the factor reversal test.

Time reversal test is a test to determine whether a given method will work both ways in time, forward and backward. In the words of Fisher, "*The test is that the formula for calculating the index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base.*" In other words, when the data for any two years are treated by the same method, but with the bases reversed, the two index numbers secured should be reciprocals of each other so that their product is unity. Symbolically, the following relation should be satisfied:

$$P_{01} \times P_{10} = 1$$

where P_{01} is the index for time "1" on time "0" as base and P_{10} is the index for time "0" on time "1" as base. If the product is not unity, there is said to be a time bias in the method. Thus, if from 2009 to 2010 the price of wheat increased from Rs. 2220 to Rs. 2960 per quintal, the price in 2010 should be 133.33 per cent of the price in 2009 and the price in 2009 should be 75 per cent of the price in 2010. One figure is the reciprocal of the other; their product (1.333 × 0.75) is unity.

The test is not satisfied by Laspeyres method and the Paasche method as can be seen below :

When Laspeyres method is used :

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0}; \text{ and } P_{10} = \frac{\sum p_0 q_1}{\sum p_1 q_1}$$

$$P_{01} \times P_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1 \text{ and the test is not satisfied.}$$

When Paasche method is used :

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0}; \text{ and } P_{10} = \frac{\sum p_0 q_0}{\sum p_1 q_0}$$

$$P_{01} \times P_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1 \text{ and the test is not satisfied.}$$

There are five methods which do satisfy the test:

(1) The Fisher's Ideal formula.

(2) Simple geometric mean of price relatives.

(3) Aggregate with fixed weights.

(4) The weighted geometric mean of price relatives with fixed weights.

(5) Marshall-Edgeworth method.

Let us now see how Fisher's Ideal formula satisfies the test.

According to Fisher's Method:

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

Changing time, *i.e.*, 0 to 1 and 1 to 0

$$P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = \sqrt{1} = 1.$$

Since $P_{01} \times P_{10} = 1$, the Fisher's Ideal index satisfies the test.

2. Factor Reversal Test

Another test suggested by Fisher is known as factor reversal test. It holds that the product of price index and the quantity index should be equal to the corresponding value index. In the words of Fisher, *"Just as each formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results, i.e., the*

two results multiplied together should give the true value ratio." In other words, the test is that the change in price multiplied by the change in quantity should be equal to the total value of given commodity in a given year is the product of the quantity and the price per unit (total value = $p \times q$). The ratio of the total value in one year to the total value in the preceding year is $\frac{p_1q_1}{p_1q_0}$. From one year to the next, both price and quantity should double, the price relative would be 200, the quantity relative 200, and the value relative 400. The total value in the second year would be four times the value in the first year. In other words, if p_1 and p_0 represent prices and q_1 and q_0 the quantities in the current year and the base year, respectively, and if P_{01} represents the change in price in the current year and Q_{01} the change in quantity in the current year, then :

$$P_{10} \times Q_{01} = \frac{\sum p_1q_1}{\sum p_0q_0}$$

If the product is not equal to the value ratio, there is, with reference to this test, an error in one or both of the index numbers.

The factor reversal test is satisfied *only* by the Fisher's Ideal Index.

Proof.

$$P_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}}$$

Changing p to q and q to p

$$Q_{01} = \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1} \times \frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}} = \sqrt{\frac{(\sum p_1q_1)^2}{(\sum p_0q_0)^2}} = \frac{\sum p_1q_1}{\sum p_0q_0}$$

Since $P_{01} \times Q_{01} = \frac{\sum p_1q_1}{\sum p_0q_0}$ the factor reversal test is satisfied by the Fisher's ideal index.

This means, of course, that the formula serves equally well for constructing indices of quantities as for constructing indices of prices, the quantity index being derived by interchanging p and q in the ideal formula. None of the simple or weighted forms of elementary indices—arithmetic mean, harmonic mean, geometric mean—fulfil the requirements of factor reversal test. It is thus obvious that the strong restrictions imposed by the factor reversal test compel its being ignored in the construction of many highly reputable index numbers.

Some of the authorities on the subject argue that there are no logical reasons for claiming that an index number ought to meet these tests. For example, Karmel has pointed out that as far as time reversal test is concerned collection of goods included in P_{01} is different from that included in P_{10} (q_0 as against q_1) and therefore, one could hardly hope for consistent results.

3. Circular Test

Another test of the adequacy of index number formula is what is known as 'circular test'. If in the use of index numbers interest attaches not merely to a comparison of two years, but to the measurement of price changes over a period of years, it is frequently desirable to shift the base. A formula is said to meet this test if, for example, the 2010 index with 2000 as the base is 200, and the 2000 index with 1995 as the base is again 200, then the 2010 index with 1995 as the base must be 400. Clearly, the desirability of this property is that it enables us to adjust the index values from period to period without referring each time to the original base. A test of this shiftability of base is called the circular test.

This test is just an extension of the time reversal test. The test requires that if an index is constructed for the year a on base year b, and for the year b on base year c, we ought to get the same result as if we calculated direct an index for a on base year c without going through b as an intermediary.

Symbolically, if there are three years a, b, c the circular test will be satisfied if :

$$\frac{P_b}{P_a} \times \frac{P_c}{P_b} \times \frac{P_a}{P_c} = 1$$

The Laspeyres index does not satisfy the test as can be seen from the following:
If the three years are 0, 1, 2; the index by the Laspeyres method will be

$$\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_2 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_2}{\sum p_2 q_2}.$$

The product of all these is not equal to 1. Hence the test is not satisfied. Similarly, it can be shown that the Paasche index and Fisher's index do not satisfy the test. However, the simple aggregative method and the fixed weight aggregative method satisfy the test as can be seen from the following :
When test is applied to the simple aggregative method, we will get

$$\frac{\sum p_1}{\sum p_0} \times \frac{\sum p_2}{\sum p_1} \times \frac{\sum p_0}{\sum p_2} = 1.$$

Similarly, when applied to fixed weight aggregative method we will get :

$$\frac{\sum p_1 q}{\sum p_0 q} \times \frac{\sum p_2 q}{\sum p_1 q} \times \frac{\sum p_0 q}{\sum p_2 q} = 1.$$

The circular test (which amounts, in fact, to a modification of the time reversal test) is met when

$$P_{ba} \times P_{cb} \times P_{ca} = 1.$$

An index which satisfies this test has the advantage of reducing the computation every time a change in the base year has to be made. Such index numbers can be adjusted from year to year without referring each time to the original base.

The circular test is not met by the ideal index or by any of the weighted aggregative with changing weights. The test is met by *simple geometric mean of price relatives and the weighted aggregative fixed weights.* The reason why Laspeyres' and Paasche's index numbers and their derivatives, the Marshall-Edgeworth and the Ideal indices, do not meet the circular test is that the weights in these index numbers depend on the periods between which comparisons are being made. If these periods change, the weights change. For example, if the base period is taken as period 2 rather than period 0, the weights in Laspeyres' index are no longer q_0 but q_2.

Karmel has pointed out that although it may seem reasonable to argue that if a price index between periods 0 and 1 has risen to M and between periods 1 and 2 to N, then between periods 0 and 2 it should have risen to MN. A moment's reflection will show that this requirement is not reasonable. An index number has meaning only in terms of the system of weighting adopted, and one may produce many numerically different but quite valid indices for comparing two periods. The weighting system used in P_{02} (Laspeyres) is the same as that in P_{01} (Laspeyres), but different from in P_{12} (Laspeyres). Consequently, the increase in M is an increase in something different from that in which N is the increase. The product MN is, therefore, a mixture, the exact meaning of which is not clear and which could not be expected to equal a direct comparison between periods 0 and 2.

Illustration 6. Show with the help of the following data that the Time and Factor Reversal Tests are satisfied by Fisher's Ideal Formula for index number construction.

Commodity	*Base Year Price (Rs.)*	*Base Year Quantity (kg.)*	*Current Year Price (Rs.)*	*Current Year Quantity (kg.)*
A	6	50	10	56
B	2	100	2	120
C	4	60	6	61
D	8.5	30	12	24
E	8	40	16	22

Solution : COMPUTATIONS FOR TIME REVERSAL TEST AND FACTOR REVERSAL TEST

Commodity	*Base year price (Rs.)* p_0	*Base year quantity (kg.)* q_0	*Current year price (Rs.)* p_1	*Current year quantity (kg.)* q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	6	50	10	56	500	300	560	336
B	2	100	2	120	200	200	240	240
C	4	60	6	61	360	240	366	244
D	8.5	30	12	24	360	255	288	204
E	8	40	16	22	640	320	352	176
					$\Sigma p_1q_0 = 2,060$	$\Sigma p_0q_0 = 1,315$	$\Sigma p_1q_1 = 1,806$	$\Sigma p_0q_1 = 1,200$

Time reversal test is satisfied when : $P_{01} \times P_{10} = 1$

$$P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \text{ and } P_{10} = \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}}$$

Substituting the values $P_{01} \times P_{10} = \sqrt{\frac{1,900}{1,360} \times \frac{1,880}{1,344} \times \frac{1,344}{1,880} \times \frac{1,360}{1,900}} = \sqrt{1} = 1.$

Hence time reversal test is satisfied.

Factor reversal test is satisfied when : $P_{01} \times Q_{01} = \frac{\Sigma p_1q_1}{\Sigma p_0q_0}$

$$P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \text{ and } Q_{01} = \sqrt{\frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times \frac{\Sigma q_1p_1}{\Sigma q_0p_1}}$$

Hence, $P_{01} \times Q_{01} = \sqrt{\frac{1,900}{1,360} \times \frac{1,880}{1,344} \times \frac{1,344}{1,360} \times \frac{1,880}{1,900}}$

$$= \frac{1,880}{1,360} \text{ which is also the value of } \frac{\Sigma p_1q_1}{\Sigma p_0q_0}$$

Hence Fisher's Ideal Index satisfies the factor reversal test.

THE CHAIN INDEX NUMBERS

In the fixed base method discussed so far the base remains the same throughout the series of the index. This method, though convenient, has certain limitations. As time elapses conditions which were once important become less significant and it becomes more difficult to compare accurately present conditions with those of a remote period. New items may have to be included and old ones may have to be deleted in order to make the index more representative. In such cases it may be desirable to use the chain base method. When this method is used the comparisons are not made with a fixed base : rather the base changes from year to year. For example, for 2009, 2008 will be the base ; for 2010, 2009 will be the base and so on. If, however, it is desired to associate these relatives to a common base, the results may be chained to obtain chain indices. Thus in its simplest form, the chain index is one in which the figures for each year (or sub-period thereof) are first expressed as percentages of the preceding year. These percentages are then chained together by successive multiplication to form a chain index.

Steps in Constructing Chain Index

(*i*) Express the figures for each year as percentages of the preceding year. The results so obtained are called link relatives.

(*ii*) Chain together these percentages by successive multiplication to form a chain index. Chain index of any year is the average link relative of that year multiplied by chain index of previous year divided by 100. In the form of formula :

$$\text{Chain Index} = \frac{\text{Current year link relative} \times \text{Previous year chain index}}{100}$$

The link relatives obtained in step (*i*) facilitate comparison from one year to another, *i.e.*, between closely situated periods in which the q's are not likely to have changed much. The chain indices obtained in step (*ii*) by a process of chaining binary comparisons facilitate long-term comparisons.

Chain relatives differ from fixed-base relatives in computation. Chain relatives are computed from link relatives whereas fixed based relatives are computed directly from the original data. The results obtained by the two different methods should be the same, but they may differ from each other slightly due to rounding off of decimal places. Since the process of computing chain relatives is quite complicated and the results are same as the fixed-base relatives obtained from the original data, chain relatives should be used when the original data are not available but the link relatives are.

Illustration 7. From the following data of the wholesale prices of wheat for the ten years, construct index numbers taking 2001 as base. Also find the link relatives.

Year	*Price of wheat (Rs. per 10 kg.)*	*Year*	*Price of wheat (Rs. per 10 kg.)*
2001	50	2006	78
2002	60	2007	82
2003	62	2008	84
2004	65	2009	88
2005	70	2010	90

Solution : CONSTRUCTION OF INDEX NUMBERS TAKING 2001 AS BASE

Year	*Price of wheat*	*Index Number (2001 = 100)*	*Year*	*Price of wheat*	*Index Number (2001 = 100)*
2001	50	100	2006	78	$\frac{78}{50} \times 100 = 156$
2002	60	$\frac{60}{50} \times 100 = 120$	2007	82	$\frac{82}{50} \times 100 = 164$
2003	62	$\frac{62}{50} \times 100 = 124$	2008	84	$\frac{84}{50} \times 100 = 168$
2004	65	$\frac{65}{50} \times 100 = 130$	2009	88	$\frac{88}{50} \times 100 = 176$
2005	70	$\frac{70}{50} \times 100 = 140$	2010	90	$\frac{90}{50} \times 100 = 180$

This means that from 2001 to 2002 there was a 20 per cent increase; from 2001 to 2003 there was 24 per cent increase. If we are interested in finding out increase from 2001 to 2002, from 2002 to 2003, from 2003 to 2004, we shall have to compute the chain indices.

CALCULATION OF LINK RELATIVES

Year	*Price of wheat*	*Link relatives*
2001	50	100.0
2002	60	$\frac{60}{50} \times 100 = 120.0$
2003	62	$\frac{62}{60} \times 100 = 103.3$
2004	65	$\frac{65}{62} \times 100 = 104.8$
2005	70	$\frac{70}{65} \times 100 = 107.7$

2006	78	$\frac{78}{70} \times 100 = 111.4$
2007	82	$\frac{82}{78} \times 100 = 105.1$
2008	84	$\frac{84}{82} \times 100 = 102.4$
2009	88	$\frac{88}{84} \times 100 = 104.8$
2010	90	$\frac{90}{88} \times 100 = 102.3$

Illustration 8. Calculate the fixed base index number and chain base index numbers from the following data. Are the two results same? If not, why ?

Commodity	*Price (in Rs. thousand)*				
	2006	*2007*	*2008*	*2009*	*2010*
I	2	3	5	7	8
II	8	10	12	14	18
III	4	5	7	9	12

Solution. Since base year is not specified the first year in order of time, *i.e.*, 2006 is taken as base. As no weights are given the appropriate method for calculating fixed base numbers is the price relatives method.

FIXED BASE INDEX NUMBERS

Commodity	*Price (in Rs. thousand)*				
	2006	*2007*	*2008*	*2009*	*2010*
I	100	$\frac{3}{2} \times 100 = 150$	$\frac{5}{2} \times 100 = 250$	$\frac{7}{2} \times 100 = 350$	$\frac{8}{2} \times 100 = 400$
II	100	$\frac{10}{8} \times 100 = 125$	$\frac{12}{8} \times 100 = 150$	$\frac{14}{8} \times 100 = 175$	$\frac{18}{8} \times 100 = 225$
III	100	$\frac{5}{4} \times 100 = 125$	$\frac{7}{4} \times 100 = 175$	$\frac{9}{4} \times 100 = 225$	$\frac{12}{4} \times 100 = 300$
Total	300	400	575	750	925
Average, *i.e.*, fixed base I.No.	100	133.3	191.7	250.0	308.3

CHAIN BASE INDEX NUMBERS CHAINED TO 2006

Commodity	*Percentage based on preceding year*				
	2006	*2007*	*2008*	*2009*	*2010*
I	100	150	166.7	140.00	114.3
II	100	125	120.0	116.67	128.6
III	100	125	140.0	128.60	133.3
Total of Link Relatives	300	400	426.7	385.3	376.2
Average	100	133.33	142.23	128.43	125.40
Chain indices	100	133.33	189.64	240.55	305.41

On comparison we find that except for first two years, the series of index numbers obtained by fixed base and chain base method are different. It is because when fixed base and chain base index numbers are computed by combining two or more series chain index numbers will be usually different from fixed base index number except for the first two given years.

Conversion of Chain Index to Fixed Base Index

At times, it may be desired to convert chain base index numbers (C.B.I.) into two fixed base index numbers (F.B.I). The following formula is used for this purpose :

$$\text{Current years' F.B.I.} = \frac{\text{Current year's C.B.I.} \times \text{Previous year's F.B.I.}}{100}$$

The following example shall illustrate the procedure :

Illustration 9. From the chain base index numbers given below prepare fixed base index numbers :

Year	2005	2006	2007	2008	2009	2010
Chain Base Index	80	105	102	95	110	120

Solution. CONVERTING CHAIN BASE INDICES TO FIXED BASE INDICES

Year	*Chain Base Index*	*Conversion*	*Fixed Base Index*
2005	80	—	80.0
2006	105	$\frac{105 \times 80}{100}$	84.0
2007	102	$\frac{102 \times 84}{100}$	85.7
2008	95	$\frac{95 \times 85.7}{100}$	81.4
2009	110	$\frac{110 \times 81.4}{100}$	89.5
2010	120	$\frac{120 \times 89.5}{100}$	107.4

Merits and Demerits of the Chain Base Method

Merits. 1. The chain base method has a great significance in practice because in business data we are more often concerned with making comparisons with the previous period and not with any distant past. The link relatives obtained by chain base method serve this purpose.

2. Chain base method permits the introduction of new commodities and the deletion of old ones without necessitating either the recalculation of entire series or other drastic changes. Thus account may readily be taken of basic changes in production, distribution and consumption habits, changes in quality, etc. Because of this flexibility chain index is used in many types of indices such as the consumer price index and the wholesale price index.

3. Weights can be adjusted as frequently as possible. This flexibility is of great significance in many types of index numbers.

4. Index numbers calculated by the chain base method are free to a greater extent from seasonal variations than those obtained by the other method.

However, one *drawback* of the chain index is that while the percentage of previous year figures give accurate comparisons of year-to-year changes, the long-range comparisons of chained percentages are not strictly valid. However, when the index number user wishes to make year-to-year comparisons, as is so often done by the businessman, the percentages of the preceding year provide a flexible and useful tool.

BASE SHIFTING, SPLICING AND DEFLATING THE INDEX NUMBERS

Base Shifting

One of the most frequent operation necessary in the use of index number is changing the base of an index. Such a change is usually referred to as shifting the base. There may be two reasons for this :

1. The previous base has become too old and is almost useless for purposes of comparison. In practice, it is desirable that the base period chosen for comparison purposes be a period of economic stability which is not too far distant in the past.

2. Comparison is to be made with another series of index numbers having a different base. For example, the consumer price index for a certain region is available with 2004 as base (*i.e.*, 2004 = 100). Now suppose an investigator wants to compare cost of living changes in the community with those of another region for which the corresponding index is given with the base year 2010. In such a case, it shall be necessary to shift the base of the first series from 2004 to 2010.

When base period is to be changed, one possibility is to recompute all index numbers using the new base period. A simpler approximate method is to divide all index numbers for the various years corresponding to the old base period by the index number corresponding to the new base period, expressing the results as percentages. These results represent the new index numbers, the index number for the new base period being 100.

Mathematically speaking, this method is strictly applicable only if the index numbers satisfy the circular test. However, for many types of index numbers the method, fortunately, yields results which in practice are close enough to those which would be obtained theoretically.

Illustration 10. The following index numbers of prices (2001 = 100) are given :

Year	*Index*	*Year*	*Index*
2001	100	2006	410
2002	110	2007	400
2003	120	2008	380
2004	200	2009	370
2005	400	2010	340

Shift the base from 2001 to 2010 and recast the index numbers.

Solution. INDEX NUMBERS WITH 2001 AS BASE (2001 = 100)

Year	*Index Numbers (2001 = 100)*	*Index Numbers (2007 = 100)*	*Year*	*Index Numbers (2001 = 100)*	*Index Numbers (2007 = 100)*
2001	100	$\frac{100}{400} \times 100 = 25.0$	2006	410	$\frac{410}{400} \times 100 = 102.5$
2002	110	$\frac{110}{400} \times 100 = 27.5$	2007	400	$\frac{400}{400} \times 100 = 100.00$
2003	120	$\frac{120}{400} \times 100 = 30.0$	2008	380	$\frac{380}{400} \times 100 = 95.00$
2004	200	$\frac{200}{400} \times 100 = 50.0$	2009	370	$\frac{370}{400} \times 100 = 92.50$
2005	400	$\frac{400}{400} \times 100 = 100.0$	2010	340	$\frac{340}{400} \times 100 = 85.00$

The new series with 2007 as base is obtained easily by dividing each entry of the first column by 400, *i.e.*, the index for 2007 and multiplying the ratio by 100.

Thus

$$\text{Index number for 2001} = \frac{\text{Index number for 2001}}{\text{Index number for 2007}} \times 100 = \frac{100}{400} \times 100 = 25.0$$

$$\text{Index number for 2002} = \frac{\text{Index number for 2002}}{\text{Index number for 2007}} \times 100 = \frac{110}{400} \times 100 = 27.5$$

In a similar manner other indices can also be obtained.

It should be carefully noted that the above method of shifting the base will not necessarily coincide with the method in which we start a new with the original data and recompute the whole series with the new base. It all depends on how the index is constructed and what weights are being used. Nevertheless,

since it is sometimes impossible to do otherwise in practice, the simple method illustrated above is often employed regardless of whether a complete recomputation of the index would produce the identical results.

Splicing

The problem of combining two or more overlapping series of index numbers into one continuous series is called splicing. The need for splicing arises for securing continuity in comparison. It happens quite often that an index is discontinued because its base has become too old. A new index may be started with same items and some recent year as base. If it is desired to connect the new index number with that of one discontinued the second number would be spliced to the first one with the result that the index would enable comparison with the old base. The process of splicing is very simple and akin to that used in shifting the base as can be seen from the following illustration.

Illustration 11. The index A given was started in 2001 and continued up to 2006, in which year another index *B* was started. Splice the index *B* to index *A* so that a continuous series of index number from 2001 up to-date may be available.

Year	*Index A*	*Index B*	*Year*	*Index A*	*Index B*
2001	100		2006	138	
2002	110		2007	150	100
2003	112		2008		120
—			2009		140
—			2010		130

Solution. INDEX B SPLICED TO INDEX A

Year	*Index A*	*Index B*	*Index B spliced to Index A 2001 as base*
2001	100		
2002	110		
2003	112		
—			
—			
2006	138		
2007	150	100	$\frac{150}{100} \times 100 = 150$
2008		120	$\frac{150}{100} \times 120 = 180$
2009		140	$\frac{150}{100} \times 140 = 210$
2010		130	$\frac{150}{100} \times 130 = 195$

The spliced index now refers to 2001 as base and we can make a continuous comparison of index numbers from 2001 onwards.

In the above case, it is also possible to splice the new index in such a manner that a comparison could be made with 2001 as base. This would be done by multiplying the old index by the ratio $\frac{100}{150}$. Thus, the spliced index for 2001 would be $\frac{100}{150} \times 100 = 66.7$, for 2002, $\frac{110}{150} \times 100 = 73.3$ for 2003, $\frac{112}{150} \times 100 = 74.7$, etc. This process appears to be more useful because a recent year can be kept as a base. However, much would depend upon the object.

It shall be clear from above that splicing is very useful for enabling comparisons between new and old index numbers. However, it should be noted that splicing can give accurate results only where geometric mean has been used in constructing the index numbers becuase in such a case index numbers are reversible. However, because of difficulties of computation, the geometric mean is not very often used in constructing index numbers.

Use of Index Numbers in Deflating

By deflating we mean making allowances for the effect of changing price levels. A rise in price level means a reduction in the purchasing power of money. To take the case of a single commodity, suppose the price of wheat rises from Rs. 1500 per quintal in 2005 to Rs. 3000 per quintal in 2010, it means that in 2010 one can buy 50 kg. of wheat for Rs. 1500 which he was spending on wheat in 2005 or, in other words, the value of rupee is only 50 paise in 2010 as compared to 2005. Thus, the value (or purchasing power) of a rupee is simply the reciprocal of an appropriate price index written as proportion. If prices increase by 60% the price index is 1.60 and what a rupee will buy is only 1/1.60 or 5/8 of what it used to buy. In other words, the purchasing power of the rupee is 5/8 of what it was or approximately 63 paise. Similarly, if prices increase by 25 per cent the price index is 1.25 (125 per cent), and the purchasing power of the rupee is 1/1.25 = 0.80 = 80 paise.

It shall be clear from above that since the value of money goes down with the rising price the workers or the salaried people are interested not so much in money wages as in real wages, *i.e.*, not how much they earn but how much their income or wage will buy.

For calculating real wages we can multiply money wages by a quantity measuring the purchasing power of the rupee, or better we divide the cash wages by an appropriate price index. This process is referred to as deflating. In principle, it appears to be very simple but in practice the main difficulty consists in finding appropriate index to deflate a given set of values or appropriate deflators. The process of deflating can be expressed in the form of formula as:

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Price Index}} \times 100$$

$$\text{Real Wage Index No.} = \frac{\text{Index of Money Wage}}{\text{Price Index}}$$

Illustration 12. Following table gives the weekly wages of workers together with the Price Index Numbers. Compute the Index numbers of real income and interpret them.

Year	*Weekly Wages (inRs.)*	*Price Index*	*Year*	*Weekly Wages (in Rs.)*	*Price Index*
2004	300	100	2008	480	350
2005	340	160	2009	570	420
2006	450	280	2010	575	430
2007	460	290			

Solution : INDEX NUMBER OF REAL INCOME

Year	*Weekly Wages (in Rs.)*	*Price Index*	*Real Wages*	*Real Wage Indices (2004 = 100)*
2004	300	100	$\frac{300}{100} \times 100 = 300.00$	100.00
2005	340	160	$\frac{340}{160} \times 100 = 212.50$	70.83
2006	450	280	$\frac{450}{280} \times 100 = 160.71$	53.57

2007	460	290	$\frac{460}{290} \times 100 = 158.62$	52.87
2008	480	350	$\frac{480}{350} \times 100 = 137.14$	45.71
2009	570	420	$\frac{570}{420} \times 100 = 135.71$	45.24
2010	575	430	$\frac{575}{430} \times 100 = 133.72$	44.57

The index number of real wages has fallen from 100 in 2004 to 44.57 in 2010. In other words, despite the fact that the weekly wage has increased from Rs. 300 in 2004 to Rs. 575 in 2010, the workers are not better off.

The method discussed above is frequently used to deflate individual values, value series or value indices. Its special use is in problems dealing with such diversified things as rupee sales, rupee inventories of manufacturers, wholesalers and retailers' incomes, wages and the like.

CONSUMER PRICE INDEX NUMBERS

Meaning and Need

The consumer price index numbers, also known as cost of living index numbers, are generally intended to represent the average change over time in the prices paid by the ultimate consumer of a specified basket of goods and services. The need for constructing consumer price indices arises because the general index numbers fail to give an exact idea of the effect of the change in the general price level on the cost of living of different classes of people, since a given change in the level of prices affects different classes of people in different manners. Different classes of people consume different types of commodities and even these same type of commodities are not consumed in the same proportion by different classes of people. For example, the consumption pattern of rich, poor and middle class people varies widely. Not only this, the consumption habits of the people of the same class differ from place to place. For example, the mode of expenditure of a lower division clerk living in Delhi may differ widely from that of another clerk of the same category living in, say, Chennai. The consumer price index helps us in determining the effect of rise and fall in prices of different classes of consumers living in different areas. The construction of such an index is of great significance because very often the demand for a higher wage is based on the cost of living index and the wages and salaries in most countries are adjusted in accordance with the consumer price index.

It should be carefully noted that the cost of living index does not measure the actual cost of living nor the fluctuations in the cost of living due to causes other than the change in the price level; its objects is to find out how much the consumers of a particular class have to pay more for a certain basketful of goods and services in a given period compared to the base period. To bring out clearly this fact, the Sixth International Conference of Labour Statisticians recommended that the term 'cost of living index' should be replaced in appropriate circumstances by the terms '*price of living index*', '*cost of living price index*'. At present, the three terms, namely, cost of living index, consumer price index and retail price index, are used in different countries with practically no difference in their connotation.

It should be clearly understood at the very outset that two different indices representing two different geographical areas cannot be used to compare actual living costs of the two areas. A higher index for one area than for another with same period is no indication that living costs are higher in the one than in the other. All it means is that as compared with the base period, prices have risen in one area than in another. But actual costs depend not only on the rise in prices as compared with the base period, but also on the actual cost of living for the base period which will vary for different regions and for different classes of population.

Utility of the Consumer Price Indices

The Consumer Price Indices are of great significance as can be seen from the following :

(1) The most common use of these indices is in wage negotiations and wage contracts. Automatic adjustments of wage or dearness allowance component of wages are governed in many countries by such indices.

(2) At Government level, the index numbers are used for wage policy, price policy, rent control, taxation and general economic policies.

(3) The index numbers are also used to measure changing purchasing power of the currency, real income, etc.

(4) Index numbers are also used for analysing markets for particular kinds of goods and services.

Construction of a Consumer Price Index

The following are the steps in constructing a consumer price index :

(1) *Decision about the class of people for whom the index is meant.* It is absolutely essential to decide clearly the class of people for whom the index is meant, *i.e.,* whether it relates to industrial workers, teachers, officers, etc. The scope of the index must be clearly defined. For example, when we talk of teachers we are referring to primary teachers, middle class teachers, etc., or to all the teachers taken together. Along with the class of people it is also necessary to decide the geographical area covered by the index. Thus in the example taken above it is to be decided whether all the teachers living in Delhi are to be included or those living in a particular locality of Delhi, say, Chandni Chowk area, Karol Bagh, etc.

(2) *Conducting family budget enquiry.* Once the scope of the index is clearly defined the next step is to conduct a family budget enquiry covering the population group for whom the index is to be designed. The object of conducting a family budget enquiry is to determine the amount that an average family of the group included in the index spends on different items of consumption. While conducting such an enquiry, therefore, the quantities of commodities consumed and their prices are taken into account. The consumption pattern can thus be easily ascertained. It is necessary that the family budget enquiry amongst the class of people to whom the index series is applicable should be conducted during the base period. The Sixth International Conference of Labour Statisticians held in Geneva in 1946 suggested that the period of enquiry of the family budget and the base period should be identical as far as possible.

The enquiry is conducted on a random basis. By applying lottery method some families are selected from the total number and their family budgets are scrutinized in detail. The items on which the money is spent are classified into certain well accepted groups, namely :

(*i*) Food,

(*ii*) Clothing,

(*iii*) Fuel and Lighting,

(*iv*) House Rent, and

(*v*) Miscellaneous.

Each of these groups is further divided into sub-groups. For example, the broad group 'food' may be divided into wheat, rice, pulses, sugar, etc. The commodities included are those which are generally consumed by people for whom the index is meant. Through family budget enquiry an average budget is prepared which is the standard budget for that class of people. While constructing the index only such commodities should be included as are not subject to wide variations in quality or to wide seasonal alternations in supply and for which regular and comparable quotations of prices can be obtained.

(3) *Obtaining price quotations.* The collection of retail prices is a very important and, at the same time, very tedious and difficult task because such prices may vary from place to place, shop to shop and person to person. Price quotations should be obtained from the localities in which the class of people concerned reside or from where they usually make their purchases. Some of the principles recommended to be observed in the collection of retail price data required for purposes of construction of cost of living indices are described below :

(*a*) The retail price should relate to a fixed list of items and for each item, the quality should be fixed by means of suitable specifications.

(*b*) Retail prices should be those actually charged to consumer for cash sales.

(*c*) Discount should be taken into account if it is automatically given to all customers.

(*d*) In a period of price control or rationing, where illegal prices are charged openly, such prices should be taken into account along with the control prices.

The most difficult problem in practice is to follow principle (*a*), *i.e.,* the problem of keeping the weights assigned and qualities of the basket of goods and services constant with a view to ensuring that only the effect of price change is measured. To conform to uniform qualities, the accepted method is to draw up detailed descriptions or specifications of the items priced for the use of persons furnishing or collecting the price quotations.

Since prices form the most important component of cost of living indices considerable attention has to be paid to the methods of price collection and to the price collection personnel. Prices are collected usually by special agents or through mailed questionnaire or in some cases through published price lists. The greatest reliance can be placed on the price collection through special agents as they visit the selected retail outlets and collect the prices from them. However, these agents should be properly selected and trained and should be given a manual of instructions as well as manual of specifications of items to be priced. Appropriate methods of price verification should be followed such as '*check pricing*' in which price quotations are verified by means of duplicate prices obtained by different agents or '*purchase checking*' in which actual purchases of goods are made.

After quotations have been collected from all retail outlets, an average price for each of the items included in the index has to be worked out. Such averages are first calculated for the base period of the index and later every month if the index is maintained on a monthly basis. The month of averaging the quotations should be such as to yield unbiased estimates of average prices as being paid by the group as a whole. This, of course, will depend upon the method of selection of retail outlets and also the scope of the index.

In order to convert the prices into index numbers the prices or their relatives must be weighted. The need for weighting arises because the relative importance of various items for different classes of people is not the same. For this reason, the cost of living index is always a weighted index. While conducting the family budget enquiry the amount spent on each commodity by an average family is decided and these constitute the weights. Percentages of expenditure on the different items constitute the 'individual *weights' allocated to the corresponding price relative and the percentage expenditure on the five groups* constitute the 'group weight'.

Methods of Constructing the Index

After the above mentioned problems are carefully decided the index may be constructed by applying any of the following methods :

(1) Aggregate Expenditure method or Aggregative methods, and

(2) Family Budget method or the method of Weighted Relatives.

1. Aggregate Expenditure Method. When this method is applied, the quantities of commodities consumed by the particular group in the base year are estimated which constitute the weights. The prices of commodities for various groups for the current year are multiplied by the quantities consumed in the base year and the aggregate expenditure incurred in buying those commodities is obtained. In a similar manner, the prices of the base year are multiplied by quantities of the base year and aggregate expenditure for the base period is obtained. The aggregate expenditure of the current year is divided by the aggregate expenditure of the base year and the quotient is multiplied by 100. Symbolically,

$$\text{Consumer Price Index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100.$$

This is in fact the Laspeyres method discussed earlier. This method is the most popular method for constructing consumer price index.

2. Family Budget Method. When this method is applied the family budgets of a large number of people for whom the index is meant are carefully studied and the aggregate expenditure of an average family on various items is estimated. These constitute the weights. The weights are thus the value weights obtained by multiplying the price by quantities consumed (*i.e.*, p_0q_0). The price relatives for each commodity are obtained and these price relatives are multiplied by value weights for each item and the product is divided by the sum of the weights. Symbolically,

$$\text{Consumer Price Index} = \frac{\Sigma PV}{V}$$

where,

$$P = \frac{p_1}{p_0} \times 100 \text{ for each item.}$$

V = Value weights, *i.e.*, p_0q_0.

This method is the same as the weighted average of price relatives method discussed earlier.

This should be noted that the answer obtained by applying the aggregate expenditure method and the family budget method shall be the same.*

Illustration 13. Prices per unit of the items forming consumption bundle of an average middle class family in two periods and percentage of total family budget allocated to those items are given in the following table :

	Food	*Rent*	*Clothing*	*Fuel*	*Misc.*
Per cent expenditure	35	15	20	10	20
Price (Rs.) in period 0	1500	500	1000	200	600
Price (Rs.) in period 1	1740	600	1250	250	900

Compute an appropriate index number and comment on the result.

Solution. The appropriate index number here would be the Consumer Price Index number.

CONSTRUCTION OF CONSUMER PRICE INDEX NUMBER

Index of Expenditure	p_0	p_1	$\frac{p_1}{p_0} \times 100$ P	W	PW
Food	1500	1740	116	35	4,060
Rent	500	600	120	15	1,800
Clothing	1000	1250	125	20	2,500
Fuel	200	250	125	10	1,250
Misc.	600	900	150	20	3,000
				$\Sigma W = 100$	$\Sigma PW = 12,610$

$$\text{Consumer Price Index} = \frac{\Sigma PW}{\Sigma W} = \frac{12610}{100} = 126.1$$

*The denominator and numerator in both methods are the same as can be seen from the following :

$\Sigma p_1 q_0$ of the Laspeyers' method is the same as ΣPV of the family budget method.

$PV = (p_1 q_0) \times p_0 q_0$ which is nothing but $p_1 q_0$.

The denominator in both methods is also the same.

$$\Sigma p_0 q_0 = \Sigma V$$

Thus there has been an increase of 26.1 per cent in the Consumer Price Index in the current year.

Illustration 14. Construct a consumer price index number from the table given below :

Group	*Index for 2003*	*Expenditure*
1. Food	550	46%
2. Clothing	215	10%
3. Fuel and Lighting	220	7%
4. House Rent	150	12%
5. Miscellaneous	275	25%

Solution : CONSTRUCTION OF CONSUMER PRICE INDEX NUMBER

Group	*Index number* I	*Expenditure* V	IV
Food	550	46	25,300
Clothing	215	10	2,150
Fuel and Lighting	220	7	1,540
House Rent	150	12	1,800
Miscellaneous	275	25	6,875
		$\Sigma V = 100$	$\Sigma IV = 37,665$

$$\text{Consumer Price Index} = \frac{\Sigma IV}{\Sigma V} = \frac{37665}{100} = 376.65.$$

Precautions while Using Consumer Price Index

Quite often consumer price indices are misinterpreted. Hence while using these indices the following points should be kept in mind :

1. As pointed out earlier, the consumer price index measure changes in the retail prices only in the given period compared to base period—it does not tell us anything about variation in the living standards at two different places. Thus if the cost of living index for working class for Mumbai is 175 and for Delhi 150 for the same period and for the same class of people it does not necessarily mean that living costs are higher in Mumbai compared to Delhi.

2. While constructing the index it is assumed that the quantities of the base year are constant and hold good for current year also. But this assumption does not appear to be very logical because the pattern of consumption goes on changing with change in fashion, introduction of new commodities in the market, etc. It is desirable, therefore, that while constructing the index the current quantities are taken into account. But this is a difficult task. The Sixth International Conference of Labour Statisticians recommended that the pattern of consumption should be examined and the weights adjusted, if necessary, at intervals of not more than ten years to correspond changes in the consumption pattern. The index also does not take into account changes in qualities. Unlike changes in consumption pattern changes in qualities of goods and services are more frequent and when a marked change in the quality of items occurs appropriate adjustments should be made to ensure that the index takes into account change in qualities also. But in practice it is a difficult proposition to follow, and, therefore, constant qualities are assumed at two different dates which again is a shaky assumption.

3. Like any other index the consumer price index is based on a sample. While constructing the index sampling is used at every stage—in the selection of commodities, in obtaining price quotations, selecting families for family budget enquiry, etc. The accuracy of index thus hinges upon the use of sampling methods. The consumption pattern derived upon the use of sampling methods. The consumption pattern derived from the expenditure data of a sample of households covered in

the course of family budget enquiry has to be representative of all the items in the average budget, the localities from which price data are collected have to be representative of all localities from which the population group makes purchases, the retail outlets from which prices are collected have to be representative of all the retail outlets patronised by the population group, etc. However, it is often difficult to ensure perfect representativeness and in the absence of this the index may fail to provide the real picture.

INDEX NUMBER OF INDUSTRIAL PRODUCTION

The index number of industrial production is designed to measure increase or decrease in the level of industrial production in a given period compared to some base period. It should be noted that such an index measures change in the *quantum* of production and not in *values*. For constructing such an index it is necessary to obtain data about the level of industrial output in the base period and the given period. Usually data about production are collected under the following heads :

1. Textile Industries—cotton, woollen, silk, etc.
2. Mining Industries—iron ore, coal, copper, petroleum, etc.
3. Metallurgical Industries—iron and steel, etc.
4. Mechanical Industries—locomotives, ships, aeroplanes, etc.
5. Industries subject to excise duties—sugar, tobacco, match, etc.
6. Miscellaneous—glass, soap, chemical, cement, etc.

The figures of output for the various industries classified above are obtained on a monthly, quarterly or yearly basis. Weights are assigned to various industries on the basis of some criteria such as capital invested, turnover, net output, production, etc. Usually the weights in the index are based on the values of net output of different industries. The index of industrial production is obtained by taking the simple arithmetic mean or geometric mean of the relatives. When simple arithmetic mean is used, the formula for constructing the index becomes :

$$\text{Index of industrial production} = \frac{\Sigma\left(\frac{q_1}{q_0}\right)W}{\Sigma W}$$

where q_1 = Quantity produced in the given period
q_0 = Quantity produced in the base period
W = Relative importance of different items.

For determining the relative share of an individual output to total output the concept of value added is most commonly used.

Illustration 15. Construct the index number of business activity in India from the following data :

Item	*Weightage*	*Index*
Industrial production	36	250
Mineral production	7	135
Internal trade	24	200
Financial activity	20	135
Exports and imports	7	325
Shipping activity	6	300

Solution. CONSTRUCTION OF INDEX NUMBER OF BUSINESS ACTIVITY

Item	*Weightage* W	*Index* I	IW
Industrial production	36	250	9,000
Mineral production	7	135	945
Internal trade	24	200	4,800
Financial activity	20	135	2,700
Exports and imports	7	325	2,275
Shipping activity	6	300	1,800
	$\Sigma W = 100$		$\Sigma IW = 21,520$

$$\text{Index No. of Business Activity} = \frac{\Sigma IW}{\Sigma W} = \frac{21,520}{100} = 215.2.$$

Limitations of Index Numbers

Though the index numbers are of great significance, the reader must also be aware of their limitations so that he avoids errors of interpretation. The chief limitations of index numbers are

1. Since index numbers are generally based on a sample, it is not possible to take into account each and every item in the construction of the index.

2. While taking the sample, random sampling is seldom used. This is so because to take sample from thousands of commodities and services, the random procedure could be neither practical nor representative. Typically, indices are constructed from samples deliberately selected. This introduces errors and every effort must be made to minimise these errors.

3. It is often difficult to take into account changes in the quality of products. With the passage of time, tastes and habits of people also change with the result that very often old commodities go out of use and new commodities are introduced. In a really typical index, qualities of commodities should remain the same over a period of time because differences in quality would mean differences in prices also. But very often it is not practicable and it makes comparisons over long period less reliable.

4. A large number of methods are designed for constructing index numbers and different methods of computation give different results. Very often the selection of an appropriate formula creates problem and in the interest of comparability, it is necessary to ensure that the same formula is adopted over a period of time for constructing a particular index. There is no such method of constructing index number which is best from every point of view. Index numbers are specialised averages and are subject to the same limitations as that of average.

5. Just like other statistical tools, index numbers can also be manipulated in such a manner as to draw the desired conclusions. Choosing a freak year is a favourite trick of those who use statistics to mislead. A dishonest capitalist could choose a record year of profits as base and so 'prove' subsequent profits to be pitifully low. Similarly, in order to prove that the current prices are intolerably high dishonest trade unionist may choose a year of exceptionally low prices as base.

6. Since in the construction of index numbers a large number of factual questions are involved, lack of adequate and accurate data in most cases becomes a serious limitation of the index itself. In most cases one cannot collect the data himself and therefore one has to rely on published sources. Ordinarily we draw upon many sources of data which are geographically dispersed. Problems of comparability and reliability thus multiply and the chances of spurious results are increased. One mistake may 'bias' the index such as including the price of one commodity for one time period, or the price of a slightly different commodity for another period or taking the manufacturer's price at one time and wholesale price at another time.

MISCELLANEOUS ILLUSTRATIONS

Illustration 16. Compute Laspeyres', Paasche's and Fisher's price index number for 2010, using the following data concerning three commodities :

Commodity	2009		2010	
	Price (Rs.)	Quantity (kg.)	Price (Rs.)	Quantity (kg.)
A	15	15	22	12
B	20	5	27	4
C	4	10	7	5

Solution : CALCULATION OF VARIOUS INDICES

Commodity	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	15	15	22	12	330	225	264	180
B	20	5	27	4	135	100	108	80
C	4	10	7	5	70	40	35	20
					$\Sigma p_1q_0 = 535$	$\Sigma p_0q_0 = 365$	$\Sigma p_1q_1 = 407$	$\Sigma p_0q_1 = 280$

Laspeyres' Index : $P_{01} = \frac{\Sigma p_1q_1}{\Sigma p_0q_0} \times 100 = \frac{535}{365} \times 100 = 146.58$

Paasche's Index : $P_{01} = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100 = \frac{407}{280} \times 100 = 145.36$

Fisher's Index : $\sqrt{L \times P} = \sqrt{146.58 \times 145.36} = 145.97$

Illustration 17. Calculate the index from the following data using Fisher's Ideal formula :

Commodity	2009 Base Year		2010 Current Year	
	Price	Quantity	Price	Quantity
A	10	50	12	60
B	8	30	9	32
C	5	35	7	40

Solution. COMPUTATION OF FISHER'S IDEAL INDEX

Commodity	2009 Base Year		2010 Current Year		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	10	50	12	60	500	600	600	720
B	8	30	9	32	240	270	256	288
C	5	35	7	40	175	245	200	280
				Total	915	1,115	1,056	1,288

Using Fisher's ideal index formula,

$$P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100 = \sqrt{\frac{1115}{915} \times \frac{1288}{1056}} \times 100$$

$$= \sqrt{1.4863} \times 100 = 1.2191 \times 100 = 121.91.$$

Illustration 18. The following table gives the weekly wages (in Rs.) of a worker and the general index number of prices during 2002-2010. Prepare the index number to show the changes in the real weekly wages of the worker.

Year	Weekly Wages (Rs.)	Price Index No.	Year	Weekly Wages (Rs.)	Price Index No.
2002	360	100	2007	640	290
2003	420	104	2008	680	300
2004	500	115	2009	720	320
2005	550	160	2010	750	
2006	600	280			

Solution : INDEX NUMBER SHOWING CHANGES IN THE REAL WEEKLY WAGES OF THE WORKER

Year	*Weekly Wages (in Rs.)*	*Price Index*	*Real Weekly Wages*	*Real Weekly Wages Indices No.*
2002	360	100	$\frac{360}{100} \times 100 = 360.00$	100.00
2003	420	104	$\frac{420}{104} \times 100 = 403.85$	112.18
2004	500	115	$\frac{500}{115} \times 100 = 434.78$	120.77
2005	550	160	$\frac{550}{160} \times 100 = 343.75$	95.49
2006	600	280	$\frac{600}{280} \times 100 = 214.29$	59.53
2007	640	290	$\frac{640}{290} \times 100 = 220.69$	61.30
2008	680	300	$\frac{680}{300} \times 100 = 226.67$	62.96
2009	720	320	$\frac{720}{320} \times 100 = 225.00$	62.50
2010	750	330	$\frac{750}{330} \times 100 = 227.27$	63.13

Illustration 19. In 2009 for working class people, wheat was selling at an average price of Rs. 160 per 10 kg., cloth at Rs. 40 per metre, house rent Rs. 10,000 per house and other items at Rs. 100 per unit. By 2010 cost of wheat rose by Rs. 40 per 10 kg., house rent by Rs. 1,500 per house and other items doubled in price. The working class cost of living index for the year 2010 (with 2009 as base) was 160. By how much the cloth rose in price during the period 2009-10 ?

Solution. Let the rise in price of cloth be X.

INDEX NUMBER FOR 2010

Commodity	*Price*	*Index No.*	*Price 2010*	*Index No.*
Wheat	160	100	200	$\frac{200}{160} \times 100 = 125$
Cloth	40	100	X	$\frac{X}{40} \times 100 = 2.5X$
House rent	10,000	100	11,500	$\frac{11,500}{10,000} \times 100 = 115$
Miscellaneous	100	100	200	$\frac{200}{100} \times 100 = 200$
				$440 + 2.5X$

The index for 2010 as given is 160. Therefore, the sum of the index number of the four commodities would be $160 \times 4 = 640$.

Hence, $440 + 2.5X = 640$

$2.5X = 200$ or $X = 80$

Hence the rise in the price of cloth was Rs. 40 (80 – 40) per metre.

Illustration 20. Owing to change in prices the consumer price index of the working class in a certain area rose in a month by one quarter of what it was before to 225. The index of food became 252 from 198, that of clothing from 185 to 205, that of fuel and lighting from 175 to 195 and that of miscellaneous from 138 to 212. The index of rent, however, remained unchanged at 150. It was known that the weight of clothing, rent and fuel and lighting were the same. Find out the exact weight of all the groups.

(MBA, Delhi Univ., 2005)

Solution. Suppose the weights of the groups are as follows :
Food X.
Fuel and Lighting Z.
Miscellaneous Y.
Rent Z.
Clothing Z.
Therefore, the index weighted index in the beginning of the month would be :

	Index I	*Weight* W	IW
Food	198	X	$198X$
Clothing	185	Z	$185Z$
Fuel and Lighting	175	Z	$175Z$
Rent	150	Z	$150Z$
Miscellaneous	138	Y	$138Y$
			$X+Y+3Z = 198X+138Y+510Z$

$$\therefore \text{ Index number} = \frac{198X+138Y+510Z}{X+Y+3Z}$$

Similarly, the weighted index at the end of the month would be :

	I	W	IW
Food	252	X	$252X$
Clothing	205	Z	$205Z$
Fuel and Lighting	195	Z	$195Z$
Rent	150	Z	$150Z$
Miscellaneous	212	Y	$212Y$
			$X+Y+3Z = 252X+212Y+550Z$

$$\text{Index number} = \frac{252X+212Y+550Z}{X+Y+3Z}$$

The weighted index at the end of the month was 225 (given). This index is a rise from the first index by one quarter. Therefore, the index at the beginning was 4/5 of 225=180.

Hence the weighted index at the beginning of the month was

$$180 = \frac{198X+138Y+510Z}{X+Y+3Z}$$

$$180X+180Y+540Z = 198X+138Y+510Z$$

or $$18X-42Y-30Z = 0 \quad ...(i)$$

Similarly, the weighted index at the end of month was

$$225 = \frac{252X+212Y+550Z}{X+Y+3Z}$$

$$225X+225Y+675Z = 252X+212Y+550Z$$

or $$27X-13Y-125Z = 0 \quad ...(ii)$$

Let the total weight be equal to 100.

Hence $$X+Y+3Z = 100 \quad ...(iii)$$

Multiplying Eqn. (*iii*) by 18 and subtracting from (*i*), we get

$$-60Y-84Z = -1800$$

or $$60Y+84Z = 1800 \quad ...(iv)$$

Multiplying (*iii*) by 27, and subtracting from eqn. (*ii*), we get

$$-40Y-206Z = -2700$$

or $$40Y+206Z = 2700 \quad ...(v)$$

Multiplying Eqn. (*iv*) by 20, and Eq. (*v*) by 30, and subtracting, we get

$$-4500Z = -45000 \text{ or } Z = 10$$

Substituting the value of Z in Eqn. (*iv*)

$$60Y + (84 \times 10) = 1800$$

$$60Y = 1800 - 840 = 960 \text{ or } Y = 16$$

Substituting the value of Y and Z in Eqn. (*iii*)

$$X + 16 + (3 \times 10) = 100$$

$$X = 100 - 16 - 30 = 54$$

∴ Thus the exact weights are :

Food	54
Clothing	10
Fuel and Lighting	10
Rent	10
Miscellaneous	16

Illustration 21. The subgroup indices of the consumer price index number for urban non-manual employees of an industrial centre for a particular year (with base 2005 = 100) were :

Food	200
Clothing	130
Fuel and Lighting	120
Rent	150
Miscellaneous	140

The weights are 60, 8, 7, 10 and 15 respectively. It is proposed to fix dearness allowance in such a way as to compensate fully the rise in the prices of food and house rent.

What should be the dearness allowance expressed as a percentage of wage ?

Solution. Let the income of the consumer be 100 rupees. He spent 60 rupees on food and 10 rupees on house rent in 2005. Since the index of food is 200 and the house rent 150 for the particular year for which the data are given, in order to maintain the same consumption standards regarding two items, he will have to spend Rs. 120 on food and Rs. 15 on house rent. Since the weights of other items are constant, in order to maintain the same standard he will have to spend 120 + 8 + 7 + 15 + 5 = Rs. 155. Hence the dearness allowance should be 55 per cent.

Illustration 22. Given the data :

	Commodities	
	A	*B*
p_0	1	1
q_0	10	5
p_1	2	X
q_1	5	2

where p and q respectively stand for price and quantity and subscripts stand for time period. Find X, if the ratio between Laspeyres' (L) and Paasche's (P) Index number is :

$$L : P = 28 : 27$$

Solution. Calculate Laspeyres' and Paasche's Indices and equate them to the given ratio in order to determine the value of X.

Commodities	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	1	10	2	5	20	10	10	5
B	1	5	X	2	$5X$	5	$2X$	2
					$\Sigma p_1q_0 = 20 + 5X$	$\Sigma p_0q_0 = 15$	$\Sigma p_1q_1 = 10 + 2X$	$\Sigma p_0q_1 = 7$

Laspeyres' Index:

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} = \frac{20 + 5X}{15}$$

Paasche's Index :

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} = \frac{10 + 2X}{7}$$

$$\frac{\frac{20+5X}{15}}{\frac{10+2X}{7}} = \frac{28}{27} \quad \text{or} \quad \frac{20+5X}{15} \times \frac{7}{10+2X} = \frac{28}{27}$$

$$\frac{140+35X}{150+30X} = \frac{28}{27}$$

$$4200 + 840X = 3780 + 945X$$

$$105X = 420 \text{ or } X = 4$$

Note : In order to work with the ratio, 100 has been omitted from the formula.

Illustration 23. An increase of 50% in the cost of a certain consumption article raises the cost of living of a certain family by 5%. What percentage of its cost of living was due to buying that article before the change in the price ?

Solution. Let the cost of the article before rise be 'X'. After increase, it, therefore, was $\frac{150X}{100} = 1.5X$, *i.e.*, the rise was $1.5X - X = 0.5X$ which is equal to an increase of 5% in the cost of living of which we call 'Y', *i.e.*, 'Y', after the increase it became $\frac{105Y}{100} = 1.05Y$ or, in other words, the increase was $1.05Y - Y = 0.05Y$.

Hence $0.5X = 0.05Y$

$$X = \frac{0.5Y}{0.5} = 0.1Y = 10\% \text{ of } Y$$

Thus the expenditure on that item was 10% of the cost of living.

Illustration 24. Compute index number from the following data using Fisher's ideal index formula :

Commodity	*Base Year*		*Current Year*	
	Qty.	*Price*	*Qty.*	*Price*
A	12	10	15	12
B	15	7	20	5
C	24	5	20	9
D	5	16	5	14

(MBA, M.D. Univ., 2005)

Solution : CALCULATION OF FISHER'S IDEAL INDEX

Commodity	*Base year*		*Current year*					
	q_0	p_0	q_1	p_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	12	10	15	12	144	120	180	150
B	15	7	20	5	75	105	100	140
C	24	5	20	9	216	120	180	100
D	5	16	5	14	70	80	70	80
					$\Sigma p_1q_0 = 505$	$\Sigma p_0q_0 = 425$	$\Sigma p_1q_1 = 530$	$\Sigma p_0q_1 = 470$

$$P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$$

$$= \sqrt{\frac{505}{425} \times \frac{530}{470}} \times 100 = \sqrt{1.3399} \times 100 = 1.1576 \times 100 = 115.76$$

Illustration 25. Compute the chain index numbers with 2006 prices as base, from the following table giving the average wholesale prices of the commodities *A*, *B* and *C* for the years 2006 to 2010.

Average wholesale price (Rs.)

Commodity	*2006*	*2007*	*2008*	*2009*	*2010*
A	20	16	28	35	21
B	25	30	24	36	45
C	20	25	30	24	30

Solution : COMPUTATION OF CHAIN INDICES

Commodity	*2006*	*Relatives based on preceding year* 2007	2008	2009	2010
A	100	$\frac{16}{20} \times 100 = 80$	$\frac{28}{16} \times 100 = 175$	$\frac{35}{28} \times 100 = 125$	$\frac{21}{35} \times 100 = 60$
B	100	$\frac{30}{25} \times 100 = 120$	$\frac{24}{30} \times 100 = 80$	$\frac{36}{24} \times 100 = 150$	$\frac{45}{36} \times 100 = 125$
C	100	$\frac{25}{20} \times 100 = 125$	$\frac{30}{25} \times 100 = 120$	$\frac{24}{30} \times 100 = 80$	$\frac{30}{24} \times 100 = 125$
Average of link relatives	300	325	375	355	310
Average of relatives	100	108.33	125	118.33	103.33
Chain index	100	$\frac{108.33 \times 100}{100}$	$\frac{125 \times 108.33}{100}$	$\frac{118.33 \times 135.41}{100}$	$\frac{103.33 \times 160.23}{100}$
		= 108.33	=135.41	= 160.23	= 165.57

Illustration 26. Construct the cost of living index number from the following group data :

Group	*Weights*	*Group Index No.*
Food	47	247
Fuel & Lighting	7	293
Clothing	8	289
House Rent	13	100
Misc.	14	236

(MBA, Vikram Univ., 2005)

Solution. CALCULATION OF COST OF LIVING INDEX

Group	*Weights* W	*Group Index* I	IW
Food	47	247	11609
Fuel and Lighting	7	293	2051
Clothing	8	289	2312
House Rent	13	100	1300
Misc.	14	236	3304
	$\Sigma W = 89$		$\Sigma IW = 20576$

$$\text{Cost of living index} = \frac{\Sigma IW}{\Sigma W} = \frac{20576}{89} = 231.19.$$

Illustration 27. The table below shows the average wages in rupees of a group of industrial workers during the year 1999 to 2010. The consumer price indices for these years with 1999 as base are also shown :

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Average wage of workers	1119	1133	1144	1157	1175	1184	1189	1194	1197	1213	1228	1245
Consumer Price index (1999 = 100)	100	107.6	106.6	107.6	116.2	118.8	119.8	120.2	119.9	121.7	125.9	129.3

(*a*) Determine the real wages of the workers during the years 1999-2010 as compared with their wages in 1999.

(*b*) Determine the purchasing power of the rupee for the year 2010 as compared to the year 1999. What is the significance of this result ?

Solution. (*i*) For finding the real wages we have to divide average wage of workers by the consumer price index.

Year	*Average wage of workers*	*Consumer Price Index (1999 = 100)*	*Real wages*
1999	1119	100	$\frac{1119}{100} \times 100 = 1119$
2000	1133	107.6	$\frac{1133}{107.6} \times 100 = 1053$
2001	1144	106.6	$\frac{1144}{106.6} \times 100 = 1073$
2002	1157	107.6	$\frac{1157}{107.6} \times 100 = 1075$
2003	1175	116.2	$\frac{1175}{116.2} \times 100 = 1011$
2004	1184	118.8	$\frac{1184}{118.8} \times 100 = 997$
2005	1189	119.8	$\frac{1189}{119.8} \times 100 = 992$
2006	1194	120.2	$\frac{1194}{120.2} \times 100 = 993$
2007	1197	119.9	$\frac{1197}{119.9} \times 100 = 998$
2008	1213	121.7	$\frac{1213}{121.7} \times 100 = 997$
2009	1228	125.9	$\frac{1228}{125.9} \times 100 = 975$
2010	1245	129.3	$\frac{1245}{129.3} \times 100 = 963$

If we divide Re. 1 by the price index of 2010, we get the purchasing power of rupee in 1999. Thus the purchasing power of rupee in 2010 shall be 100/129.3 = 0.77. This means that the purchasing power of rupee has gone down—in 2010 the rupee could buy only 77 per cent of what it could buy in 1999.

Illustration 28. Calculate Laspeyres' and Paasche's price and quantity indices from the data given below :

Commodity	*2009*		*2010*	
	Price	*Qty.*	*Price*	*Qty.*
A	4	10	5	12
B	6	8	7	10
C	10	5	12	4
D	3	12	4	15
E	5	7	5	8

Solution. CALCULATION OF LASPEYRES' AND PAASCHE'S PRICE AND QUANTITY INDICES

Commodity	*2009*		*2010*					
	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	4	10	5	12	50	40	60	48
B	6	8	7	10	56	48	70	60
C	10	5	12	4	60	50	48	40
D	3	12	4	15	48	36	60	45
E	5	7	5	8	35	35	40	40
					$\Sigma p_1q_0 = 249$	$\Sigma p_0q_0 = 209$	$\Sigma p_1q_1 = 278$	$\Sigma p_0q_1 = 233$

Laspeyres' index

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{249}{209} \times 100 = 119.14$$

$$Q_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{233}{209} \times 100 = 111.48$$

Paasche's index

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{278}{233} \times 100 = 119.31$$

$$Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{278}{249} \times 100 = 111.65$$

Illustration 29. An enquiry into the budgets of middle class families in a village near Hyderabad gave the following information :

Expenses on :	*Food*	*Rent*	*Clothing*	*Education*	*Misc.*
	30%	25%	15%	10%	20%
Price (Rs.) 2009	1800	1000	700	400	700
Price (Rs.) 2010	2000	1200	900	500	1000

Construct cost of living index and comment on the change in the cost of living in 2010 as compared to 2009.

Solution. CONSTRUCTION OF COST OF LIVING INDEX

Expenses on	*2009* p_0	*2010* p_1	$\frac{p_1}{p_0} \times 100$ P	W	PW
Food	1800	2000	111.11	30	3333.30
Rent	1000	1200	120.00	25	3000.00
Clothing	700	900	128.57	15	1928.55
Education	400	500	125.00	10	1250.00
Misc.	700	1000	142.86	20	2857.20
				$\Sigma W = 100$	$\Sigma PW = 12369.05$

$$\text{Cost of Living Index (for 2010)} = \frac{\Sigma PW}{\Sigma W} = \frac{12369.05}{100} = 123.69$$

Illustration 30. For the following data, calculate price index number of 2010 with 2009 as the base year, using : (*a*) Laspeyres' method, (*b*) Fisher's method.

	2009		*2010*	
	Price	*Quantity*	*Price*	*Quantity*
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

Solution. CALCULATION OF PRICE INDEX NUMBERS

Commodity	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	20	8	40	6	320	160	240	120
B	50	10	60	5	600	500	300	250
C	40	15	50	15	750	600	750	600
D	20	20	20	25	400	400	500	500
					$\Sigma p_1q_0 = 2070$	$\Sigma p_0q_0 = 1660$	$\Sigma p_1q_1 = 1790$	$\Sigma p_0q_1 = 1470$

Laspeyres' Index :

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{2070}{1660} \times 100 = 124.7$$

Paasche's Index :

$$P_{10} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{1790}{1470} \times 100 = 121.77$$

Illustration 31. *Calculate from the following data, the Fisher's Ideal Index Number for the year 2010 :*

Commodity Selected	*2009*		*2010*	
	Price (Rs.)	*Expenditure on quantity consumed (Rs.)*	*Price (Rs.)*	*Expenditure on quantity consumed (Rs.)*
A	8	200	65	1950
B	20	1400	30	1650
C	5	80	20	900
D	10	360	15	300
E	27	2160	10	600

Solution : First find quantity by dividing expenditure by price.

CALCULATION OF FISHER'S IDEAL INDEX

Commodity	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	8	25	65	30	1625	200	1950	240
B	20	70	30	55	2100	1400	1650	1100
C	5	16	20	45	320	80	900	225
D	10	36	15	20	540	360	300	200
E	27	80	10	60	800	2160	600	1620
					$\Sigma p_1q_0 = 5385$	$\Sigma p_0q_0 = 4200$	$\Sigma p_1q_1 = 5400$	$\Sigma p_0q_1 = 3385$

$$P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$$

$$= \sqrt{\frac{5385}{4200} \times \frac{5400}{3385}} \times 100 = 1.430 \times 100 = 143.$$

Illustration 32. Construct Fisher's Ideal Index from the following data and show that it satisfies time reversal and factor reversal tests :

Commodity	*2009*		*2010*	
	Price	*Value*	*Price*	*Value*
A	10	100	12	144
B	15	75	20	120
C	8	80	10	110
D	20	60	25	50
E	50	500	60	540

Solution.

CALCULATION OF FISHER'S IDEAL INDEX

Commodity	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	10	10	12	12	120	100	144	120
B	15	5	20	6	100	75	120	90
C	8	10	10	11	100	80	110	88
D	20	3	25	2	75	60	50	40
E	50	10	60	9	600	500	540	450
					$\Sigma p_1q_0 = 995$	$\Sigma p_0q_0 = 815$	$\Sigma p_1q_1 = 964$	$\Sigma p_0q_1 = 788$

Fisher's Ideal Index or $$P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$$

$$= \sqrt{\frac{995}{815} \times \frac{964}{788}} \times 100 = \sqrt{1.4935} \times 100 = 1.222 \times 100 = 122.2$$

Time Reversal Test : Time reversal test is satisfied when :

$$P_{01} \times P_{10} = 1$$

$$P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{\frac{788}{964} \times \frac{815}{995}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{995}{815} \times \frac{964}{788} \times \frac{788}{964} \times \frac{815}{995}} = \sqrt{1} = 1$$

Hence time reversal test is satisfied.

Factor Reversal Test : Factor reversal test is satisfied when :

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma p_0 q_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} = \sqrt{\frac{788}{815} \times \frac{964}{995}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{995}{815} \times \frac{964}{788} \times \frac{788}{815} \times \frac{964}{995}} = \frac{964}{815}$$

$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$ is also equal to $\frac{964}{815}$. Hence factor reversal test is satisfied by the given data.

Illustration 33. From the data given below, calculate Fisher's Ideal Index and show that it satisfies time reversal test :

	2009		*2010*	
Commodity	*Price*	*Quantity*	*Price*	*Quantity*
A	12	20	14	30
B	14	13	20	15
C	10	12	15	20
D	6	8	4	10
E	8	5	6	5

Solution. CALCULATION OF FISHER'S IDEAL INDEX

	2009		*2010*					
Commodity	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	12	20	14	30	280	240	420	360
B	14	13	20	15	260	182	300	210
C	10	12	15	20	180	120	300	200
D	6	8	4	10	32	48	40	60
E	8	5	6	5	30	40	30	40
					$\Sigma p_1q_0 = 782$	$\Sigma p_0q_0 = 630$	$\Sigma p_1q_1 = 1090$	$\Sigma p_0q_1 = 870$

Fisher's Ideal Index or $$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

$$= \sqrt{\frac{782}{630} \times \frac{1090}{870}} \times 100 = 1.247 \times 100 = 124.7$$

Time Reversal Test : Time Reversal Test is satisfied when

$$P_{01} \times P_{10} = 1$$

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{\frac{870}{1090} \times \frac{630}{782}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{782}{630} \times \frac{1090}{870} \times \frac{870}{1090} \times \frac{630}{782}} = \sqrt{1} = 1.$$

Illustration 34. Calculate the index number by : (*i*) Paasche's method, and (*ii*) Fisher's method.

Commodity	p_1	q_1	p_0	q_0
A	5	14	3	8
B	8	18	6	25
C	3	25	1	40
D	15	36	12	48
E	9	14	7	18
F	7	13	5	19

(MBA, M.D. Univ., 2006)

Solution. CALCULATION OF PAASCHE'S AND FISHER'S INDICES

Commodity	p_1	q_1	p_0	q_0	p_1q_1	p_0q_1	p_1q_0	p_0q_0
A	5	14	3	8	70	42	40	24
B	8	18	6	25	144	108	200	150
C	3	25	1	40	75	25	120	40
D	15	36	12	48	540	432	720	576
E	9	14	7	18	126	98	162	126
F	7	13	5	19	91	65	133	95
					$\Sigma p_1q_1 = 1046$	$\Sigma p_0q_1 = 770$	$\Sigma p_1q_0 = 1375$	$\Sigma p_0q_0 = 1011$

(*i*) Paasche's Index : $P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{1046}{770} \times 100 = 135.84$

(*ii*) Fisher's Index : $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$

$$= \sqrt{\frac{1375}{1011} \times \frac{1046}{770}} \times 100 = 1.3592 \times 100 = 135.92$$

Illustration 35. Compute the Laspeyres' and Paasche's price index numbers for the year 2010 using the following data concerning four commodities :

	Commodity			
Quantity (*kg.*)	*A*	*B*	*C*	*D*
in 2009	8	10	15	20
in 2010	6	5	10	15
Price per kg (Rs.)				
in 2009	20	50	40	20
in 2010	40	60	50	20

Solution. CALCULATION OF LASPEYRES' AND PAASCHE'S PRICE INDEX

Commodity	p_0	p_1	q_0	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	20	40	8	6	320	160	240	120
B	50	60	10	5	600	500	300	250
C	40	50	15	10	750	600	500	400
D	20	20	20	15	400	400	300	300
					$\Sigma p_1q_0 = 2070$	$\Sigma p_0q_0 = 1660$	$\Sigma p_1q_1 = 1340$	$\Sigma p_0q_1 = 1070$

Laspeyres' Index : $$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{2070}{1660} \times 100 = 124.7$$

Paasche's Index : $$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_1} \times 100 = \frac{1340}{1070} \times 100 = 125.23$$

PROBLEMS

1-A : Answer the following questions, each question carries **one** mark:

(*i*) What are index numbers ?
(*ii*) Give two important uses of index numbers.
(*iii*) Name two important problems that arise while constructing index number.
(*iv*) Give the formula for Fisher's Ideal Index Number.
(*v*) What is time reversal test ?
(*vi*) What is Laspeyres method of constructing index numbers.
(*vii*) What is quantity index ?
(*viii*) What is base shifting ?
(*ix*) Which average is most appropriate for constructing index numbers.
(*x*) Give any two limitations of index numbers.

1-B : Answer the following questions, each question carries **four** marks:

(*i*) Describe the problem faced in the construction of index numbers. *(M.A. Econ., Madras Univ., 2003)*
(*ii*) Distinguish between Time reversal test and Factor reversal test. *(M.B.A., UPTech. Univ., 2007)*
(*iii*) What is Fisher's Ideal Index ? Why is it called ideal ?
(*iv*) Briefly explain the concept of splicing and deflating.
(*v*) What are fixed base and chain base indices ? Explain with suitable example.

2. (*a*) What is an index number ? Describe briefly its applications in business and industry.
(*b*) Discuss briefly the importance and the use of index numbers in business.

3. (*a*) What are the uses of index numbers ? What are the problems in their construction? *(MBA, Vikram Univ., 2006)*
(*b*) What are index numbers ? How are they constructed ? Explain the role of weights in the construction of general price index numbers.
(*c*) Explain the nature and uses of index numbers.

4. What is an index number ? Examine the various problems involved in the construction of an index number. Discuss in brief the uses of an index number.

5. What is an index number ? Explain the terms price relative, quantity relative and value relative with reference to a single commodity and deduce the factor reversal property.

6. Describe the steps involved in the computation of Fisher's Ideal Index Number. What are its advantages and disadvantages ?

7. What is Fisher's Ideal Index ? Why is it called ideal ? Show that it satisfies both the time reversal test as well as the factor reversal test. *(MBA, Sukhadia Univ.; MBA, HPU 2004)*

8. Laspeyres' price index generally shows an upward trend in the price changes while Paasche's method shows a downward trend on them. Elucidate the statement. *(MBA, Delhi Univ., 2005)*

9. (*a*) "Index numbers are signs and guide-posts along the business highway, that indicate to the businessman how he should drive or manage his affairs". Explain the above statement and also point out the relative advantages of the various types of averages as applied to index numbers. Which would you prefer and why ?
(*b*) What is Fisher's Index ? Why is it called Ideal ?

10. Discuss the following statements :

(*i*) Compute
(*i*) "The purpose determines the type of index number to use."
(*ii*) "An index number is a special type of average."
(*iii*) "There is no such thing as unweighted index numbers."
(*iv*) "The choice of a suitable base period is at best a temporary solution." Why ?
(*v*) "Theoretically, geometric mean is the best average in the construction of index numbers but in practice mostly arithmetic mean is used." Why ?

11. (*a*) " Since the value of the base is always 100, it does not make any difference which period is selected as the base on which to construct an index." Comment.

(b) If you are employed to construct a price index for a department store that sells thousands of items (a) how would you decide on which items to include ? (b) how would you define the price? (c) what weights would you use? (d) which formula would you select ?

(a) Define index numbers. Describe the construction of wholesale price index number elucidating the following points :
(i) Selection of commodities,
(ii) Selection of the prices and the market,
(iii) Selection of the base year,
(iv) Selection of the average,
(v) Decision on the system of weighting.

(b) What is an ideal index ? How does the Fisher formula for ideal index satisfy the following two tests :
(i) Time Reversal Test, and
(ii) Factor Reversal Test.

(a) Distinguish clearly between fixed base and chain base index number and point out their relative merits and demerits. *(B.Com., Delhi Univ., 2009)*

(b) Explain Time Reversal Test and Factor Reversal Test with the help of a suitable example. *(MBA, HPU; MBA, Osmania Univ., 2007)*

What are time reversal and factor reversal tests ? Does the following index number formula satisfy these tests ?

$$I = \sqrt{\frac{\Sigma p_x q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_x q_x}{\Sigma p_0 q_x}} \times 100$$

(a) What is an index number ? Discuss its importance in business and industry.

(b) Explain :
(i) Time reversal test,
(ii) Factor reversal test, and
(iii) Circular test as applied to index number.

What do you understand by reversibility of index numbers ? Explain time reversal and factor reversal test in this context.

(a) It is said that index numbers are a specialized type of averages. How far do you agree with this statement. Explain briefly time reversal and factor reversal tests. *(MBA, Osmania Univ., 2004)*

(b) What are the factor reversal and circular tests of consistency in the selection of an appropriate index formula ? Verify whether Fisher's Ideal Index satisfies such tests.

(c) The following are the prices of six different commodities for 2009 and 2010. Compute a price index by (a) simple aggregative method and (b) average of price relative method by using both arithmetic mean as well as geometric mean.

Commodity	*Unit*	*Price in 2009 (Rs.)*	*Price in 2010 (Rs.)*
Wheat	Quintal	1900	2200
Rice	"	1500	2000
Pulses	"	2000	3000
Ghee	1 kilo	120	122
Butter	"	130	136
Potatoes	"	11	12

Construct an appropriate index for purposes of comparison from the following data :

Commodity	*A*		*B*		*C*	
Year	*Price*	*Qty.*	*Price*	*Qty.*	*Price*	*Qty.*
2009	4	50	3	10	2	5
2010	10	40	8	8	4	4

19. The following table gives the per capita income and cost of living index number of a particular community. Deflate the per capita income by taking into account the rise in the cost of living :

Year	*Per capita income*	*Cost of Living Index No. Base*
2003	800	100
2004	900	150
2005	950	180
2006	1020	200
2007	1150	220
2008	1200	250
2009	1500	300
2010	1600	400

20. Calculate Laspeyres' and Paasche's Index number from the following data :

	Base Year		*Current Year*	
Items	*Qty.*	*Price per Kg.*	*Qty.*	*Price per Kg.*
Bread	10	Rs. 22.50	12	Rs. 25.00
Meat	8	Rs. 80.00	9	Rs. 90.00
Tea	2	Rs. 100.00	4	Rs. 120.00

21. Construct from the following data spliced index continuous with index *A* and a spliced index continuous with index *B*:

Year	*Index A*	*Index B*
2005	100	
2006	95	
2007	110	
2008	125	110
2009		105
2010		94

22. (*a*) In the construction of a certain consumer price index, the following group index numbers were found. Calculate the consumer price index by using (*i*) the weighted arithmetic mean, and (*ii*) the weighted geometric mean :

Groups	*Index*	*Weights*
Food	300	5
Fuel and Lighting	250	1
Clothing	280	1
House Rent	200	2
Miscellaneous	150	1

(*b*) In calculating a certain cost of living index number, the following weights were used : Food 15, clothing 0, rent 4, fuels and light 2, miscellaneous 1. Calculate the index for a date when the average percentage increases in prices of items in the various groups over the base period were 32, 54, 47, 74 and 58 respectively.

23. Using the following data, show that Fisher's Ideal formula satisfies the Factor Reversal Test :

Commodity	*Price Per Unit (Rs.)*		*Number of Units*	
	Base Period	*Current Period*	*Base Period*	*Current Period*
A	6	10	50	56
B	2	2	100	120
C	4	6	60	60
D	10	12	30	24
E	8	12	40	36

[139.79]

24. Using the food index and the information given below calculate the cost of living index number :

Groups	*Food*	*Clothing*	*Fuel & Light*	*House Rent*	*Miscellaneous*
Index	–	310	220	150	300
Weight	60	5	8	9	18

25. Given below are the data on prices of some consumer goods and the weights attached to the various items. Compute price index numbers for the year 2010 (Base: 2009 = 100) using (*i*) simple average, and (*ii*) weighted average of price relatives.

		Price (Rs.)		
Item	*Unit*	*2009*	*2010*	*Weight*
Wheat	kg.	10.00	11.00	2
Milk	litre	15.00	16.00	5
Sugar	kg.	16.00	17.00	8
Shoes (per pair)	Rs.	500.00	550.00	1

26. An enquiry into the budgets of the middle-class families of a certain city revealed that on an average the percentage expenses on the different groups were—Food 45, rent 15, clothing 12, fuel and light 8 and miscellaneous 20. The group index numbers for the current year as compared with a fixed base period were respectively 410, 150, 343, 248 and 285. Calculate the consumer price index number for the current year. Mr. *X* was getting Rs. 240 in the base period and Rs. 480 in the current year. State how much he ought to have received as extra allowance to maintain his former standard of living.

(MBA, HPU, 2005)

27. The following are the group index numbers and the group weights of an average working class family's budget. Construct the cost of living index number by assuming the weight :

Group	*Index number*	*Weight*
Food	152	48
Fuel and Lighting	110	6
Clothing	130	8
House rent	100	12
Miscellaneous	90	15

[129.73]

28. From the chain base index numbers given below prepare fixed base index numbers.

Year	:	2006	2007	2008	2009	2010
Chain base Index No.	:	80	140	130	110	90

[80, 112, 145.6, 160.16, 144.1]

29. From the following data prepare index number for real wages of workers :

Year	:	2005	2006	2007	2008	2009	2010
Wages (Rs.)	:	2000	2500	3110	3600	3900	4000
Index number	:	100	160	280	300	330	340

30. Calculate the Fisher's ideal index. Does this data satisfy time and factor reversal tests :

Commodity	*Price*	*Qty.*	*Price*	*Qty.*
A	5	4	4	5
B	5	10	6	10
C	8	15	8	20
D	10	8	12	6
E	2	6	2	8

[106.63, yes]

31. The following table gives the average wholesale price of five groups of commodities for the years 2006 to 2010. Compute chain base index numbers.

Commodity	*2006*	*2007*	*2008*	*2009*	*2010*
A	2	3	4	2	7
B	3	6	9	4	3
C	4	12	20	8	22
D	5	7	22	16	18
E	3	8	11	14	12

32. Compute the index numbers of prices from the following data by applying : (*a*) Laspeyres', (*b*) Paasche's (*c*) Fisher's and (*d*) Bowley's method.

	2009		*2010*	
Commodity	*Price*	*Quantity*	*Price*	*Quantity*
A	3	8	6	9
B	5	9	8	10
C	6	15	7	12
D	4	20	5	15

[(*a*) 135.98; (*b*) 140.19; (*c*) 138.07; (*d*) 138.09]

33. Prepare index numbers (2003 = 100) from the Link relatives given below :

Year :	2004	2005	2006	2007	2008	2009	2010
Link relatives :	105	75	71	105	98	90	90

34. Calculate Laspeyres', Paasche's, and Fisher's Ideal Index from the following data :

Commodity	*Price*	*Value*	*Price*	*Value*
A	10	100	8	96
B	16	96	14	98
C	12	36	10	40
D	15	60	5	25

[P_{01}=73.29; 72.96; 73.12]

35. Prepare price index numbers for 2010 with 2000 as base year from the following data by using (*i*) Laspeyres' (*ii*) Paasche's and (*iii*) Fisher's method. (Correct up to three decimal)

	Article							
	I		*II*		*III*		*IV*	
Year	*P*	*Q*	*P*	*Q*	*P*	*Q*	*P*	*Q*
2000	12.50	9	9.63	4	7.75	6	5.00	5
2010	18.75	9	7.75	6	8.80	10	10	6.507

[*P*: Price; *Q* : Quantity]

With the help of above data prove that the Time Reversal Test is satisfied by Fisher's formula, but not necessarily by the Laspeyres' and Paasche's index numbers.

36. Construct Fisher's Ideal Index No. for the following data and show that it satisfies the time reversal and factor reversal tests :

	Base Year		*Current Year*	
Commodity	*Price*	*Qty.*	*Price*	*Qty.*
A	6	30	15	40
B	5	40	10	55
C	10	25	12	20
D	4	15	3	30
E	2	50	5	28

(*MBA, Madurai-Kamaraj Univ., 2006*)

37. From the following prices of these groups of commodities for the years 2003 to 2007, find the chain base index numbers chained to 2003 :

Groups	*2003*	*2004*	*2005*	*2006*	*2007*
I	4	6	8	10	12
II	16	20	24	30	36
III	8	10	16	20	24

(*MBA, M.K. Univ., 2007*)

*****

10

Business Forecasting and Time Series Analysis

INTRODUCTION

The growing competition, rapidity of change in business activities and the trend towards automation demand that decisions in business are not based purely on guesses and hunches rather on a careful analysis of data concerning the future course of events. More time and attention should be given to the future than to the past, and the question 'what is likely to happen ?' should take precedence over 'what has happened ?' though no attempt to answer the first can be made without the facts and figures being available to answer the second.

When estimates of future conditions are made on a systematic basis, the process is referred to as "forecasting" and the figure or statement obtained is known as a "forecast". Forecasting is a service whose purpose is to offer the best available basis for management expectations of the future and to help management understand the implications for the firm's future of the alternative courses of action to them at present. In a world where the future is not known with certainty, virtually every business and economic decision rests upon a forecast of future conditions. In fact when a person assumes the responsibility of running a business he automatically takes the responsibility for attempting to forecast the future and to a very large extent his success or failure would depend upon the ability to forecast successfully the future course of events. Forecasting aims at reducing the area of uncertainty that surrounds management decision-making with respect to costs, profit, sales, production, pricing, capital investment, and so forth. If the future were known with certainty, forecasting would be unnecessary. Decisions could be made and plans formulated on a once-for-all basis, without the need for subsequent revision. But uncertainty does exist, future outcomes are rarely assured and, therefore, organised system of forecasting is necessary.

Forecasting is concerned with two main tasks : first, the determination of the best basis available for the formation of intelligent managerial expectations; and second, the handling of uncertainty about the future, so that implication of decisions become explicit. Forecasting activity can be viewed as part of the management information system. It also impinges on the control system. Forecasts are commonly applied to capital investment decisions, strategic planning, product and market planning, production planning and stock control, budgetary control and financial planning and competitive position planning. In fact, managers are forecasters and they are forecasting for much of their time. They plan production in expectation of certain levels of sales. They set prices in expectation of certain levels of wages, raw material costs, financial constraints and sales. They build warehouses in expectation of certain levels of stocks and sales. They recruit labour, buy materials, arrange finance, or plan factories in expectation of certain levels of sales and other activity. The following are main functions of forecasting :

(1) The creation of plans of action. It is impossible to evolve a worthwhile system of business control without one acceptable system of forecasting.

(2) The second general use of forecasting is to be found in monitoring the continuing progress of plans based on forecasts. Forecasts serve the function of lighthouses to shipmasters at night, reference points for course and speed requiring action/no action decisions.

(3) The forecast provides a warning system of the critical factors to be monitored regularly because they might drastically affect the performance of the plan.

It is obvious from the above that forecasts intelligently used may serve the function of both lighthouse and compass. However, the object of business forecasting is not to determine a curve or series of figures that will tell exactly what will happen, say, a year in advance, but it is to make analysis based on definite statistical data, which will enable an executive to take advantage of future conditions to a greater extent than he could do without them. In many respects the future tends to move like the past. This is a good thing, since without some element of continuity between past, present and future, there would be little possibility of successful prediction. But history is not likely to repeat itself and we would hardly expect economic conditions next year or over the next year ten years to follow a clear-cut pattern. Yet, frequently, past patterns prevail sufficiently to justify using the past as a basis for predicting the future.

In forecasting one should note that it is impossible to forecast the future precisely—there always must be some range of error allowed for in the forecast. Statistical forecasts are those in which we can use the mathematical theory of probability to measure the risks of errors in predictions.

Steps in Forecasting

Broadly speaking, the forecasting of business fluctuations consists of the following steps :

1. *Understanding why changes in the past have occurred.* One of the basic principles of statistical forecasting—indeed of all forecasting when historical data are available—is that the forecaster should use the data on past performance to get a "speedometer reading" of the current rate (of sales, say) and of how fast that rate is increasing or decreasing. The current rate and changes in the rate—"acceleration" and "deceleration"—constitute the basis of forecasting. Once they are known, various mathematical techniques can develop projections from them. If an attempt is made to forecast business fluctuations without understanding why past changes have taken place, the forecast will be purely mechanical, based solely upon the application of mathematical formulae and subject to serious error.

2. *Determining which phases of business activity must be measured.* After it is known why business fluctuations have occurred, or if there is a reasonable supposition it is necessary to measure certain phases of business activity in order to predict what changes will probably follow the present level of activity.

3. *Selecting and compiling data to be used as measuring devices.* There is an interdependent relationship between the selection of statistical data and determination of why business fluctuations occur. Statistical data cannot be selected and compiled in an intelligent manner unless there is a sufficient understanding of business fluctuations ; likewise, it is important that reasons for business fluctuations be stated in such a manner that it is possible to secure data that are related to the reasons.

4. *Analysis of data.* In this last step, the data are analysed in the light of one's understanding of the reason why changes occur. For example, if it is reasoned that a certain combination of forces will resul in a given change, the statistical part of the problem is to measure these forces and from the data available to draw conclusions on the future course of action. The methods of drawing conclusions may be calle forecasting techniques, which represent any one of a large number of analytical devices for summarisin data and drawing inferences from the summaries.

Requirements of a Good Forecasting System

A forecasting system to be instrumental in contributing to better management decision-makin needs certain conditions :

(1) It must involve the managers whose decisions are affected.
(2) Individual forecasts and group of forecasts have to be specifically relevant to the decisions being taken.
(3) The forecasts must not claim too much validity or authority.
(4) Implications of the various probable errors in the predictions for the organisations need to be thoroughly worked through so that management can evaluate the consequences of the probable range of likely outcomes.
(5) Management must at least know how badly things could go wrong if all the guesses turned out wrong.

Methods of Forecasting

There is nothing new about business forecasting. For centuries businessmen have tried to adjust themselves in such a manner as to make the best out of the future conditions. The rule-of-thumb method has been widely practised in business. It consists of deciding about the future in terms of past experience and familiarity with the problem at hand. Even today this method is very widely used in business. However, it can lead to absurd conclusions if employed by the inexperienced.

In recent years the techniques of forecasting have improved to a marked degree and are applicable to almost every sphere of business activity. Attempts are being made to make forecasting as scientific as possible. The base of scientific forecasting is statistics, *i.e.*, numerical data on business trends which many businessmen fail to acquaint themselves with. However, forecasting business change involves more than an analysis of statistical data—it also embodies the prediction of economic change such as secular trend, seasonal variation or cyclical variation and a consideration of cause and effect. To handle the increasing variety of managerial forecasting problems, many forecasting techniques have been developed in recent years. Each has its special use, and care must be taken to select the correct technique for particular application. Also before applying a method of forecasting, the following questions should be answered :

(1) What is the purpose of the forecast—how is it to be used ?
(2) What are the dynamics and components of the system for which the forecast will be made ?
(3) How important is the past in estimating the future ?

The following are some of the important methods of forecasting :

1. Historical Analogy Method,
2. Field Surveys and Opinion Poll,
3. Business Barometers,
4. Extrapolation,
5. Regression Analysis,
6. Econometric Models,
7. Lead-Lag Analysis,
8. Exponential Smoothing,
9. Input-output or end-use Analysis,
10. Time Series Analysis,

Here only a brief description of first nine methods and detailed description of the last method, which is very popularly used in practice, is made.

1. **Historical Analogy Method.** When this method of forecasting is used, the forecast in regard to a particular phenomenon is based on some analogous conditions elsewhere in the past. For example, the

forecast for demand for steel, cement, cars, etc., in India today may be based on the same analogy of demand for these products in the U.S.A. In the year, say, 2004 if it is found that the conditions now prevailing in India are very much like those that prevailed in the United States during that period.

Analogies very much help a country in determining the various stages of growth through which a country is passing before it reaches the '*take off*' stage. Through analogies one can also have an idea of the social changes such as changes in attitudes and values, norms of social life, life style, etc. Generally, we find that as a country heads towards economic advancement many of the old beliefs and values change and people start thinking in different light altogether.

This method of forecasting is considered to be a qualitative one because it is difficult to quantify most phenomena in respect to which analogies are being made. A serious limitation of this method is that a search has to be made of such a place and period in history and it may really be difficult to get exactly comparable conditions. Hence the method can only be used as a rough guide and exclusive reliance on it should not be placed.

2. **Field Surveys and Opinion Poll.** Field surveys may be conducted to obtain the necessary information which may constitute the basis for forecasting. The survey methods are used widely for forecasting demand both of the existing and new products marketed within and outside the country. Surveys can help in obtaining both qualitative and quantitative information. However, the various survey techniques are to be used with great caution so that the element of bias in responses is minimised.

The information gathered through survey methods can be discussed with various experts and other knowledgeable persons in the field and their opinion can be obtained. For example, the opinion of sales representatives, wholesalers, retailers and other intermediaries may be obtained while formulating demand projections.

It is quite likely that experts in the different fields such as production, sales, finance may have divergent views and it may be necessary for all of them to sit together, give a patient hearing to others' viewpoints, convince others and change their opinion, if necessary. A consensus view can be obtained by the use of Delphi method.

3. **Business Barometers.** Of great assistance in practical forecasting is a series that can be used as an "index" or "indicator" of the basic conditions related to the industry. The term "barometer" is also widely though loosely used in business statistics ; sometimes the term is used to mean simply an indicator of the present economic situation and sometimes it is used to designate an indicator of future conditions.

The following are some of the important business activities which aid businessmen in forecasting :

1. Gross national product,
2. Employment,
3. Wholesale prices,
4. Consumer prices,
5. Industrial production,
6. Volume of bank deposits and currency outstanding,
7. Consumer credit,
8. Disposable personal income,
9. Departmental store sales,
10. Stock prices,
11. Bond yields,

This list is by no means exhaustive, nor is the arrangement necessarily in order of importance. Several of the above series are composite average of totals—or indexes of these averages or totals. Analysis also should be made of some of the major components of these activities.

Index numbers relating to different activities in the field of production, trade, finance, etc., may also be combined into a general index of business activity. This general index refers to the general conditions of trade and industry. But the behaviour of individual industries or trades might show a different trend from that of the Composite Business Activity Index. Also, general boom or depression may be reflected in a majority of separate industries and trades, yet some industries and trades might show quite contrary tendencies. Hence, the study of general business conditions as revealed by the Composite Business Index should be supplemented by special studies of individual businesses based on separate indices. The trends indicated by barometers will guide the businessmen as to whether the stocks of goods should be increased or decreased or whether to increase investment or not, etc.

4. **Extrapolation.** Extrapolation is the simplest yet often a useful method of forecasting. In many forecasting situations the most reasonable expectation is that the variable will follow its already established path. Extrapolation relies on the relative consistency in the pattern of past movements in some time series. Strictly speaking, nothing needs to be known about causation—why the series moves as it does. But in practice the justification for extrapolation does involve the nature of the growth process being described. Extrapolation is used frequently for sales forecasts and for other estimates when "better" forecasting methods may not be justified.

Since extrapolation assumes that the variable will follow its established pattern of growth, the problem is to determine accurately the appropriate trend curve and the values of its parameters. Numerous alternative trend curves are suitable for business forecasting application. Some of the most useful ones are :

(*a*) *Arithmetic trend.* The straight-line arithmetic trend assumes that growth will be by a constant absolute amount each year.

(*b*) *Semi-log trend.* The semi-logarithmic trend assumes constant percentage increase each year. Since the annual increment is constant in logarithms, this line translates into a straight line when drawn on paper with a logarithmic vertical scale.

(*c*) *Modified exponential trend.* This curve assumes that each increment of growth will be a constant per cent (less than 100) of the previous one. The line trends generally do approach, but never quite reach a constant asymptote, which may be thought of as an upper limit.

(*d*) *Logistic curve.* The logistic curve has both an upper asymptote and a lower asymptote. It assumes a 'law of growth' involving increasing increments from an initial low value and then gradual slowing down of growth as 'maturity' is approached.

(*e*) *The Gompertz curve.* The Gompertz curve is a curve with similar properties as described above and is often used to describe growth of industrial output.

Selection of an appropriate growth curve can be guided by empirical and theoretical considerations. Empirically, it is a question of selecting the curve that best fits the past movement of the data. Theoretical matters which intervene in that logic may support a particular growth pattern. For example, population growth when there are no resources or choice restraints imply a geometric pattern of growth, as has been known since Malthus. With limited resources, however, population is sometimes thought to grow along a logistic curve. Let these theoretical notions be taken too seriously, it should be emphasised that empirical considerations may lead us quickly to a more realistic and less restrictive notion of the relevant growth curve.

5. **Regression Analysis.** The regression approach offers many valuable contributions to the solution of forecasting problems. It is the means by which we select from among the many possible or theoretically suggested relationships between variables in a complex economy. With it, one makes the jump from intuitive evaluation on the connection between two variables to precise quantified knowledge. If two variables are functionally related then a knowledge of one will make possible an estimate of the other. For example, if we know that advertising expenditure and sales are correlated then for a given advertising expenditure, we can find out the probable increase in sales or *vice versa.*

Regression analysis may involve only one predicted, or dependent, and one independent variable—simple regression, or it may involve relationships between the variable to be forecast and several independent variables—multiple regression. Statistical techniques to estimate the regression equations are often fairly complex and time-consuming, but there are many computer programs now available that estimate simple and multiple regressions quickly without much of costs involved.

There are two dangers in using regression analysis for forecasting :

(*i*) There is possibility of a mechanistic approach, accepting with little question the relationship which the calculations reveal—perhaps that with the highest r^2 — and applying it to the forecast. There are many possibilities for spurious correlation among time series as many series move together over time even where there is no conceivable connection between them.

(*ii*) If the trend of observations follows a curve, the linear regression will still fit the best straight line to the data, but any projection will be nonsense. There is little which we can do with non-linearity graphically, and computers handle nonlinear forms as routinely as linear.

6. **Econometric Models.** The term econometric refers to the application of mathematical economic theory and statistical procedures to economic data in order to verify economic theorems and to establish quantitative results in economics. An econometrician is, therefore, an economist, a statistician and a mathematician, all in one. Econometric models take the form of a set of simultaneous equations. The values of the constants in such equations are supplied by a study of statistical time series, and large number of equations may be necessary to produce an adequate model. The work of computations is greatly facilitated by electronic data processing equipment like computer, etc.

At the present time, most short-term forecasting uses only statistical methods with little quantitative information. However, in the years to come when most large companies develop and refine econometric models of their major businesses, this tool of forecasting will become more popular. But, it should be remembered that the development of an econometric model requires sufficient data so that the correct relationships can be established. Hence when data are scarce—for example, when a product is first introduced into a market—this method cannot be profitably employed.

The econometric model is, in principle, the most formal, since the forecast is based on an explicit mathematical model. The model states in detail and in quantitative terms the way in which the various aspects of the economy are interrelated. Theoretically, the model makes possible a wholly mechanical forecast because once values have been estimated for the exogenous variable, the solution of the model gives specific values for the predicted variable. But, in actual practice qualitative and quantitative forecasters have tended to come together. The 'artist' forecaster has become fully aware of the fact that he needs quantitative relationships, while the 'econometric' forecaster has learned that in some instances, quantitative relationships have to be modified by qualitative factors.

The econometric model provides the forecaster with a record of the prediction with a clear statement of the assumptions concerning exogenous variables and the solution of the model—it is often possible—or at least it is made easier—to trace and reproduce the causes for success as well as failures. One can learn just where errors were made and, hopefully, where improvements can be made. Thus, discredited hypothesis may be dropped and new ones can be substituted which ultimately will lead to better understanding of the economic system and business fluctuations.

The econometric models are not very popular in practice because it is probably neither necessary nor feasible for every business forecaster to construct his own model of the economy. The effort and costs involved in a fully-developed econometric model are well beyond most forecasting operations. Thus, most forecasters will probably rely for some time on the basic aggregate models developed at research institutes or universities. These models may be used to make predictions and to test out alternative assumptions about Government policy or the other exogenous aspects of the economy. With the help of

the models and, hopefully, sector analysis of his own industry, the business forecaster will be in a good position to augment other more familiar approaches. The better the understanding of the various forecasting methods and of their interrelationships, the better the forecasts will be.

7. Lead-lag Analysis

The Lead-lag approach attempts to determine the approximate lapse of time between the movement of one series and the movements of general business conditions. If one or more series can be found such that their turning points lead by a number of months with substantial regularity the turning points of general business in the past, it is quite logical to use these leading series to predict what is going to happen to general business activity.

The most important list of statistical indicators in modern times originated during the 1937-38 sharp business contraction. The list was prepared by the National Bureau of Economic Research (NBER). The list comprised 21 series that on their past performance, some dating as far back as 1854, promised to be fairly reliable indicators of business revival. The list was revised three times—the most recent list comprises 26 indicators of business expansion and contraction. Among the current 26 NBER statistical indicators, 12 are classified as leading series, 8 as coincident series and 6 as lagging series. Leading indicators, as pointed out by Chou, are mainly those series which are concerned with business decisions to expand or to curtail output. Time is required to work out their effects, and so they tend to move ahead of turns in business cycles. *Leading indicators* signal in advance a change in the basic performance of the economy as a whole. *Coincident indicators* are those whose movements coincide roughly with, and provide a measure of the current performance of aggregate economic activity. Hence they inform us whether the economy is currently experiencing a slowdown or not. Movements of *lagging indicators* usually follow rather than lead those of the coincident indicators. In general, lagging indicators move in directions opposite to those of the leading indicators throughout various phases of business cycles.

8. Exponential Smoothing*

The method of exponential smoothing for forecasting is an outgrowth of the recent attempts to maintain the smoothing function of moving averages without their corresponding drawbacks and limitations. Two major limitations to the use of moving averages are :

(*i*) to compute a moving average forecast it is necessary to store the last N observation values. This takes up considerable space which in many computer systems is costly ;

(*ii*) the method of moving averages gives equal weight to each of the last N observed and no weight at all to observations before period $(t - N)$; that is, the weight given to each of the last N observations is $1/N$ and 0 (zero) for any previous observations.

In principle, exponential smoothing operates in a manner analogous to moving averages by "smoothing" historical observations to eliminate randomness. The mathematical procedure for performing this smoothing, however, is somewhat different from that used in moving averages. The basic principle and the application of this device are quite simple. If we wish to forecast the value of a time series for the period $t + 1$ on the basis of information available just after period t, the forecast is best considered as a function of two components : the actual value of the series for period, t, and the forecasted value for the same period made in the previous period $t - 1$. The use of both realised and estimated values available now for predicting future values is better than the use of either alone.

The exponential smoothing models can be either single exponential smoothing model or double exponential smoothing model, the former being applicable in the absence of trend, and the latter being applicable when the time series is exhibiting some type of growth pattern.

---

* The term exponential smoothing is obtained from the weight attached to the preceding observations.

When we talk of single exponential smoothing model, the forecasted value of the series at time period t, $\hat{y}_t$, is equal to a fraction of the forecast error of the previous period $(y_{t-1} - \hat{y}_{t-1})$, plus the forecasted values of the previous period. Thus to predict the value of the time series at time $t + 1$, we use

$$y_{t+1} = a(y_t - \hat{y}_t) + \hat{y}_t$$

It should be noted that exponential smoothing, a special kind of weighted moving average, is found to be useful in short-term forecasting for inventories and sales. Exponential smoothing can also be employed for projections over long terms. Because of several computational advantages over the simple moving average, the exponential smoothing time series model is perhaps the most widely used time series forecasting technique today.

Some experts have been critical of exponential smoothing procedures on the ground that they are empirically inadequate when major policy decisions undergo change. They also object to the assumption that differing weights should be assigned to observations depending solely upon their relative recency. The exponential smoothing as a technique for prediction has very limited applications for sales items that have many wide variations in expected demand.

9. Input-Output Analysis

The input-output (I-O) technique was developed by Professor Leontief in the 1930s. However, it is only recently that it has caught the eye of big business. Input-output analysis gets its name from the type of data on which it is based. That is, the material requirements (input) and the product (output) of every economic activity in an I-O model are the raw data. The most widely familiar model is the pioneer work of Leontief for the United States.

When input-output method is used, an input-output table is made. An input-output table is a technique for determining the transactions taking place within and among different sectors of the economy as well as the magnitude of such transactions. An input-output table summarises "taking" and "giving" among and within all industries and between them and the final consumer. In this sense it is an accounting procedure for reporting all transactions. Its main usefulness is for planning purposes at the level of national economy. Several countries, mainly European, utilise input-output tables widely for planning purposes.

The input-output tables can also be constructed for single business, organisation or industry. The different divisions, departments or organisational units can then become the entries of the table among which transactions can take place and for which forecasts can be made.

Since the construction of an input-output table can be expensive and since the technological requirements among departments are not constant, its value in forecasting is rather doubtful. The value of input-output analysis lies more in the planning stages than in forecasting itself, but it is still an expensive planning device for the purposes of individual organisations.

10. Time Series Analysis

This is the most popular method of business forecasting and is discussed in detail.

Business Forecasting and Time Series Analysis

The first step in making estimates for the future consists of gathering information from the past data. In this connection, one usually deals with statistical data which are collected, observed or recorded at successive intervals of time. Such data are generally referred to as 'time series'*. Thus when we observe numerical data at different points of time, the set of observations is known as time series. For

* "A time series is a set of observations taken at specified times, usually at 'equal intervals'. Mathematically, a time series is defined by the values Y_1, Y_2.. of a variable Y at times t_1, t_2...Thus Y is a function of t, symbolised $Y = F(t)$."

example, if we observe production, sales, imports, etc., at different points of time, say, over the last 5 or 10 years, the set of observations formed shall constitute time series. Thus, in the analysis of time series, time is the most important factor because the variable is related to time which may be either year, month, week, day, hour or even minute or second.

The problem of time series analysis can best be appreciated with the help of the following example :

The following are the figures of sales (in thousand units) of a firm :

Year	*Sales of Firm A (thousand units)*	*Year*	*Sales of Firm A (thousand units)*
2003	40	2007	43
2004	32	2008	48
2005	47	2009	65
2006	41	2010	42

If we observe the above series we find that generally the sales have increased but for some years a decline is also noticed. There may be several causes responsible for increase or decrease from one period to another such as changes in tastes and habits of people, growth of population, availability of alternative products, etc. It may be very difficult to study the effect of various factors that have led either to an increase or decrease in sales. The statistician, therefore, tries to analyse the effect of the various forces under four broad heads :

(1) Changes that have occurred as a result of general tendency of the data to increase or decrease, known as 'secular variations or trend.'

(2) Changes that have taken place during a period of 12 months as a result of change in climate, weather conditions, festivals, etc. Such changes are called 'Seasonal Variations.'

(3) Changes that have taken place as a result of booms and depressions. Such changes are classified under the head 'Cyclical Variations'.

(4) Changes that have taken place as a result of such forces that could not be predicted like floods, earthquakes, famines, etc. Such changes are classified under the head 'Irregular or Erratic Variations'.

These variations are called components of time series and shall be discussed in detail.

Role of Time Series Analysis

Time Series Analysis is of great significance in business decision-making for the following reasons :

(1) *It helps in the understanding of past behaviour.* By observing data over a period of time, one can easily understand what changes have taken place in the past. Such analysis will be extremely helpful in predicting the future behaviour.

(2) *It helps in planning future operations.* Statistical techniques have been evolved which enable time series to be analysed in such a way that the influences which have determined the form of that series may be ascertained. If the regularity of occurrence of any feature over a sufficient long period could be clearly established then, within limits, prediction of probable future variations would become possible.

In fact, the greatest potential of a time series lies in predicting an unknown value of the series. From this information intelligent choices can be made concerning capital investment decisions, decisions concerning production and inventory, etc.

(3) *It helps in evaluating current accomplishments.* The actual performance can be compared with the expected performance and the cause of variation analysed. For example, if expected sales for 2010 were 20 lacs coloured T.V. sets and the actual sales were only 19 lacs; one can investigate the cause for the shortfall in achievement. Time Series Analysis will enable us to apply the scientific

procedure of "holding other things constant" as we examine one variable at a time. For example, if we know how much is the effect of seasonality on business we may devise ways and means of ironing out the seasonal influence or decreasing it by producing commodities with complementary seasons.

(4) *It facilitates comparison.* Different time series are often compared and important conclusions drawn therefrom.

However, one should not be led to believe that by time series analysis one can foretell with 100 per cent accuracy the course of future events. After all, statisticians are not foretellers. This could be possible only if the influences of the various forces which affect these series such as climate, customs and traditions, growth and decline factors and the complex forces which produce business cycles would have been regular in their operation. However, the facts of life reveal that this type of regularity does not exist. But this then does not mean that time series analysis is of no value. When such analysis is coupled with a careful examination of current business indicators, one can undoubtedly improve substantially upon guestimates (*i.e.,* estimates based upon pure guesswork) in forecasting future business conditions.

COMPONENTS OF TIME SERIES

It is customary to classify the fluctuations of a time series into four basic types of variations which account for the changes in the series over a period of time. These four types of patterns, variations, movements, or, as they are often called components or elements of time series, are :

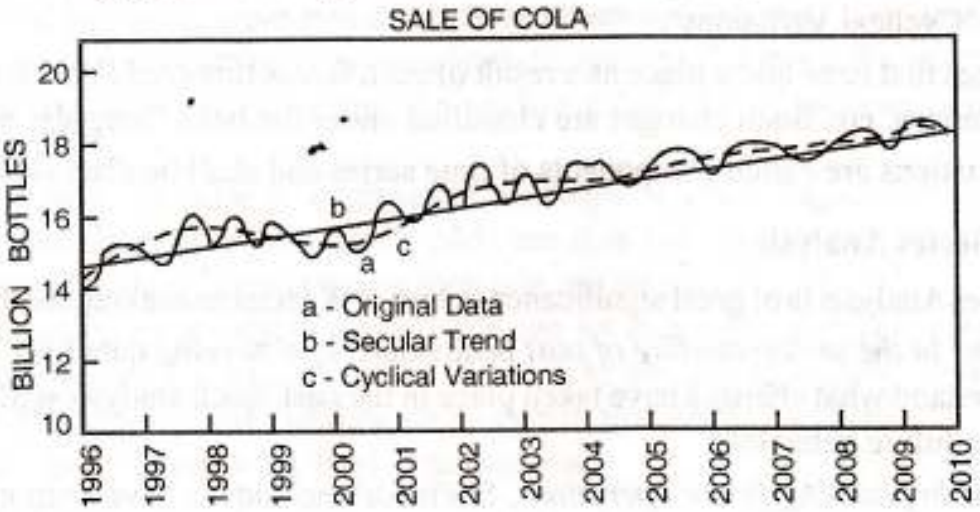
(1) Secular Trend,

(2) Seasonal Variations,

(3) Cyclical Variations,

(4) Irregular Variations.

Look at the following graph showing the sale of Cola during the years 1996 to 2010 :



The original data in this graph is represented by a curve. The general movement persisting over a long period of time represented by the diagonal line drawn through the irregular curve is called *secular trend.*

Next, if we study the irregular curve year by year, we see that in each year the curve starts with a low figure and reaches a peak about the middle of the year and then decreases again. This type of fluctuation, which completes the whole sequence of change within the span of a year and has about the same pattern year after year, is called a *seasonal variation.*

Furthermore, looking at the broken curve superimposed on the original irregular curve, we find pronounced fluctuations moving up and down every few years throughout the length of the chart. These

are known as *business cycles* or *cyclical fluctuations*. They are so called because they comprise a series of repeated sequences just as wheel goes round and round.

Finally, the little saw-tooth irregularities on the original curve represent what are referred to as *irregular movements*.

In traditional or classical time series analysis, it is ordinarily assumed that there is a multiplicative relationship between these four components, that is, it is assumed that any particular value in series is the product of factors than can be attributed to the various components. Symbolically :

$$Y = T \times S \times C \times I$$

where Y is the result of the four elements :

T = Trend,

S = Seasonal Variation,

C = Cyclical Variation, and

I = Irregular Variation.

If the above model is employed, the seasonal, cyclical and random items are not viewed as absolute amounts but rather as relative influences. Thus, a seasonal index of 110 per cent would mean that the actual value is 10 per cent higher than it otherwise would be because of seasonal influences.

This particular model is appropriate for those situations in which percentage changes best represent the movement in the series.

Another approach is to treat each observation of a time series as the sum of these four components. Symbolically,

$$Y = T + S + C + I$$

When this relationship is assumed the major aim of time series analysis is to isolate those parts of the overall variation of a time series which are traceable to each of these four components and measuring each part independently. There are numerous variations of these basic models. The models

$$Y = TCS + I \text{ or } Y = TC + SI$$

are two such variations.

There is little agreement amongst experts about the validity of the different assumptions—some feel that the given classification is too crude and that there are more than four types of movements. Nothing specific is really known about how the components are related, how they combine to produce particular effects, or whether they are really separable. The effects of the various components might be additive, multiplicative or they might be combined in any one of infinitely large number of other ways. Different models (assumptions or theories) will lead to different results. Although the additive assumption is undoubtedly true in some cases, the multiplicative assumption characterises the majority of economic time series. Consequently, *the multiplicative model is not only considered the standard or traditional assumption for series analysis, it is more often employed in practice than all other possible models combined.* For this reason, we shall discuss the multiplicative model in detail in this chapter. However, it should be kept in mind that not all time series are best represented by this model and if the examination of data reveals that certain components do not react in the prescribed fashion, then a different model may be better.

The task of performing a time series analysis is to operate on the data in such a way as to bring out separately each of the components present.

1. Secular Trend

The term 'secular trend' or simply 'trend' is very popularly used in day-to-day conversation. For example, we often talk that the population, prices, production, etc., are showing an upward trend. What

we really mean thereby is that if we observe such variables over a long period of time we find an increasing tendency. Similarly, we may find some variables showing downward tendency or constant tendency. For example, we find that over the last several years the death rate in our country is declining and hence we say that death rate is showing a declining trend. Similarly, with the improvement in the means of transport the number of bullock-carts on the road is declining year after year. However, in a dynamic economy such examples where either a downward or constant tendency is observed are rare—most of the variables show an upward trend. Thus, the general tendency of the data to grow or decline over a long period of time is technically called 'secular trend' or simply 'trend'. It should be noted that when we talk of trend, we mean thereby smooth, regular, long-term movement of the data—sudden and erratic movements either in upward or in downward direction have nothing to do with the trend.

There are all sorts of trends : some series increase slowly and some increase fast, others decrease at varying rates, some remain relatively constant for long periods of time, and some after a period of growth or decline reverse themselves and enter a period of decline or growth. Broadly speaking, the various types of trends are divided under two heads :

(1) Linear or Straight Line Trends, and

(2) Non-linear Trends.

Both these types will be discussed later in this chapter.

For a proper understanding of the meaning of trend, the reader's attention is drawn to the following two points :

(*i*) *The meaning of long term.* When we say that secular trend refers to the general tendency of the data to grow or decline over a long period of time, one may be interested in finding out as to what constitutes a long period of time. Does it mean several years ? The answer is 'no'. On the other hand, whether a particular period can be regarded as long or not in studying secular trend depends upon the nature of data. To take an example, if we are studying the figures of sales of a firm for 2009-10 and 2008-09 and we find that in 2009-10 sales have gone up, this increase cannot be called as secular trend because this is too short a period of time. On the other hand, if we put a strong germicide into bacterial culture, and count the number of organisms still alive after each 10 seconds for 8 minutes, these 48 observations showing a general pattern would be called secular movements. It is clear from this example that in one case year could not be regarded as a long period whereas in another case even 8 minutes constituted long period. Hence it depends on nature of data whether a particular period would be called as long or not.

Generally speaking, the longer the period covered, the more significant the trend. When the period is short, the secular movements cannot be expected to reveal themselves clearly and the general drift of the series may be unduly influenced by the cyclical fluctuations. This would make it difficult to separate the various series of variations in time series. *As a minimum safeguard, it may be said that to compute trend, the period should cover at least two or three complete cycles.*

(*ii*) Another point worth mentioning is that for concluding whether the data is showing an upward tendency or downward tendency it is *not necessary that the rise or fall must continue in the same direction throughout the period.* We have to observe the general tendency of the data. As long we can say that the period as a whole was characterised by an upward movement or by a downward movement, we say that a secular trend was present. For example, if we observe the trend of price over a period of 20 years and find that except for a year or two the prices are continuously rising, we would call it a secular rise in prices.

Factors Affecting Trend

There are several factors that affect trend in time series. The most important single factor responsible for rising trends in series like prices, production, sales, etc., has been the ever increasing

population. On the other hand, declining trends in certain series are the result of the technological, institutional and cultural changes, the very things which produced much of the growing trend in most of the other series. For example, the progress in automobile industry reduced on the road and, on the other hand, increased the number of cars, buses, trucks, etc. Similarly, better medical facilities, improved sanitation, diet, etc., on the one hand reduce the death rate and on the other contribute to a rise in birth rate. Such influences as these produce gradual changes. But at times it is possible that certain innovations may take place which may cause sudden changes in the outlook for particular industries or individual concerns.

The basic objective of the study of trend is to predict the future behaviour of the data. If a trend can be determined, then the rate of change or progress can be ascertained and tentative estimates concerning the future be made accordingly. For example, by projecting the trend line one can find out expected sales of a firm, for, say 201 or the expected population for 2015 or 2030 likely, and so on. Such forecasts are of immense use in framing basic policies and planning for the future. However, such forecasts are based on the assumption that the past growth has been steady and that the conditions determining this growth may reasonably be expected to persist in the future. A change in these conditions would affect the forecasts.

2. Seasonal Variations

Seasonal variations are those periodic movements in business activity which occur regularly every year and have their origin in the nature of the year itself. Since these variations repeat during a period of 12 months they can be predicted fairly accurately. Nearly every type of business activity is susceptible to seasonal influence to a greater or lesser degree and as such these variations are regarded as normal phenomenon recurring every year. Although the word 'seasonal' seems to imply a connection with the season of the year, the term is meant to include any kind of variation which is of periodic nature and whose repeating cycles are of relatively short duration. Seasonal variation is evident when the data are recorded at weekly or monthly or quarterly intervals. Although the amplitude of seasonal variations may vary, their period is fixed being one year. As a result, seasonal variations don't appear in series of annual figures. The factors that cause seasonal variations are :

(i) Climate and weather conditions. The most important factor causing seasonal variation is the climate. Changes in the climate and weather conditions such as rainfall, humidity, heat etc., act on different product and industries differently. For example, during winter there is greater demand for woollen clothes, hot drinks, etc., whereas in summer cotton clothes, cold drinks have a greater sales. Agriculture is influenced very much by the climate. The effect of the climate is that there are generally two seasons in agriculture—the growing season and harvesting season—which directly affect the income of the farmer which in turn affects the entire business activity.

(ii) Customs, traditions and habits. Though nature is primarily responsible for seasonal variations in time series, customs, traditions, and habits also have their impact. For example, on certain occasions like Deepawali, Dussehra, Christmas, etc., there is a big demand for sweets and the bank withdrawals go up because people want money for shopping and gifts, etc. Similarly, on the first of every month there are heavy withdrawals and the banks have to keep lots of cash to meet the possible demand on the basis of past experience.

The study and measurement of seasonal patterns constitute a very important part of analysis of time series. In some cases, seasonal patterns themselves are of primary concern because little, if any, intelligent planning or scheduling (or production, inventory, personnel, advertising and the like) can be done without a knowledge based on adequate statistical measures of seasonal patterns. In other cases, the seasonal variation may not be of immediate concern, but it must be measured to facilitate the study of other types

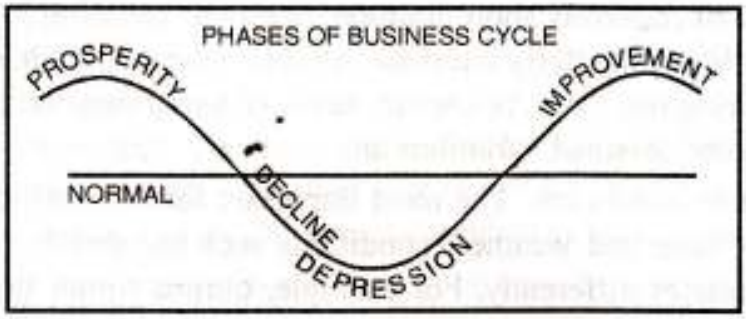
of variations based on adequate statistical measure of seasonal patterns. An accurate knowledge of seasonal behaviour is an aid in mitigating and ironing out seasonal movements through business policy. This may be done by introducing diversified products having different seasonal peaks, accumulating stock in slack seasons in order to manufacture at a more regular rate, cutting prices in slack seasons and advertising off-seasonal use for the products. Seasonal indexes are also helpful in scheduling purchases, inventory control, personnel requirement, seasonal financing and selling and advertising programmes. For example, a housewife may buy fruits for canning or preserving at the peak of the season when the prices are low and quality high. Seasonal fluctuations may also be ironed out in order that the intra-year fluctuations may be less pronounced.

3. Cyclical Variations

The term 'cycle' refers to the recurrent variations in time series that usually last longer than a year and regular, neither in amplitude nor in length.

Most of the time series relating to economics and business show some kind of cyclical variation. Cyclical fluctuations are long-term movements that represent consistently recurring rises and declines in activity. A business cycle* consists of the recurrence of the up and down movements of business activity from some sort of statistical trend or "normal". By "normal" we mean a kind of statistical average : we do not mean that there is anything very permanent or special. There are four well-defined periods or phases in the business cycle, namely : (*i*) prosperity, (*ii*) decline, (*iii*) depression, and (*iv*) improvement.

Each phase changes gradually into the phase which follows it in order given. The following diagram would illustrate a cycle. In the prosperity phase of the business cycle the public is optimistic—business is booming, prices are high and profits are easily made. There is a considerable expansion of business activity which leads to an over-development. It is then difficult to secure deliveries and there is shortage of transportation facilities, which has a tendency to cause large inventories to be



accumulated during the time of highest prices. Wages increase and labour efficiency decreases. The strong demand for money causes interest rates to rise to a high level while doubt enters the bankers mind as to the advisability of granting further loans. This situation causes businessmen to make price concessions in order to secure the necessary cash. Then follows the expectation of further reduction and the situation becomes worse instead of better. Buyers wait for lower prices and all this leads to a decline in business activity. Then follows period of pessimism in trade and industry ; factories close, businesses fail, there is widespread unemployment, while wages and prices are low. These conditions characterise the period of depression. After a period of rigid economy, liquidation and reorganisation money accumulates and

*Business cycles are a type of fluctuations found in the aggregate economic activity of nations that organize their work mainly in business enterprise, a cycle consists of expansions occurring at about the same time in many economic activities followed by similarly general recessions, contractions and revivals which merge into the expansion phase of the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years, they are not divisible into shorter cycles of similar character with amplitudes approximating their own.

—Arthur Burns and Mitchell

seeks a use. Then follows a period of increasing business activity with rising prices, a period of improvement or recovery. The improvement period generally develops into the prosperity period and a business cycle is completed. Then movements discussed above are constantly repeated in the order given as the cycle completes it swing every few years.

The study of cyclical variations is extremely useful in framing suitable policies for stabilising the level of business activity, *i.e.*, for avoiding periods of booms and depressions as both are bad for an economy—particularly depression which brings about a complete disaster and shatters the economy.

But despite the great importance of measuring cyclical variations, they are the most difficult type of economic fluctuations to measure. It is because of the following two reasons :

(*i*) Business cycles do not show regular periodicity—they differ widely in timing, amplitude and pattern which makes their study very tough and tedious.

(*ii*) The cyclical variations are mixed with erratic, random or irregular forces which make it impracticable to isolate separately the effect of cyclical and irregular forces.

Business cycles are distinguished from seasonal variations in the following respects :

(*i*) The cyclical variations are of a longer duration than a year. A business cycle may be of any duration but normally the period of business cycle is 2-10 years. Moreover, they do not ordinarily exhibit regular periodicity as successive cycles vary widely in timing, amplitude and pattern.

For example, the 23 cycles of general business in the United States between 1854 and 1949 averaged 40 months; in duration individual cycles differed greatly. The shortest period lasted only 20 months and the longest persisted for 29 months.

(*ii*) The fluctuations in a business cycle result from a different set of causes. The period of prosperity, decline, depression and improvement viewed as four phases of a business cycle are generated by factors other than weather, social customs, and those which create seasonal patterns.

4. Irregular Variations*

Irregular variations refer to such variations in business activity which do not repeat in a definite pattern. In fact, the category labelled irregular variation is really intended to include all types of variations other than those accounting for the trend, seasonal and cyclical movements. These latter three, if they are actually at work, act in such a way as to produce certain systematic effects. Irregular movements, on the other hand, are considered to be largely random. being the result of chance factors, which like those determining the fall of a coin, are wholly unpredictable.

Irregular variations are caused by such isolated special occurrences as flood, earthquakes, strikes and wars. Sudden changes in demand or very rapid technological progress may also be included in this category. By their very nature these movements are very irregular and unpredictable. Quantitatively, it is almost impossible to separate out the irregular movements and the cyclical movements. Therefore, while analysing time series, the trend and seasonal variations are measured separately and the cyclical and irregular variations are left altogether.

There are two reasons for recognising irregular movements:

(*i*) To suggest that on occasions it may be possible to explain certain moments in the data as due to specific causes and to simplify further analysis.

(*ii*) To emphasise the fact that prediction of economic conditions is always subject to degree of error owing to the unpredictable erratic influences which may enter.

---

*Irregular variations are also called 'erratic', 'random' or 'accidental' variations.

Problems of Classification

Although it is a simple matter to classify the factors affecting time series into four groups for analytical purposes, the actual application of the classification frequently presents serious problems. Seasonal variations are by no means always so uniform in amplitude and timing that their identification can be made with certainty. Consequently, the investigator often finds it hard to distinguish seasonal influences from cyclical or random factors. Long and severe cycles may, to some observors, appear to be changes in the direction of the regular trend. During the Great Depression of the 1930's, for example, many leading economists interpreted the existing conditions not as a cyclical depression but as "secular stagnation".

Another difficulty arises because the four components of time series data are not mutually independent of one another. An exceedingly severe seasonal influence may aggravate or even precipitate a change in the cyclical movement. Conversely, cyclical influence may seriously affect the seasonal. A very [illegible] rising trend virtually eliminates seasonal and cyclical variations.

Finally, the fourfold breakdown of time series data when applied to general business conditions has frequently been challenged on analytical grounds. Bratt* sees not one trend, but two: a primary trend representing the long-term growth of productive capacity, and the drift away from it which he calls secondary trend. Schumpeter developed an even more detailed breakdown by identifying three cyclical components, the 3-year Kitchin cycle, the 10-year Juglar cycle and the 50-year Kondratieff cycle**. The divergence of opinion among eminent scholars indicates clearly that the fourfold breakdown is more approximation, convenient to employ but frequently subject to modification.

Preliminary Adjustments before Analysing Time Series

Before beginning the actual work of analysing a time series it is necessary to make certain adjustments in the raw data. The adjustments are:

1. Adjustment for Calendar Variation,
2. Adjustment for Population changes,
3. Adjustment for Price Changes, and
4. Adjustment for Comparability.

1. *Calendar Variation.* A vast proportion of the important time series is available in a monthly for and it is necessary to recognise that the month is a variable time unit. The actual length of the shorte month is about 10 per cent less than that of longest, and if we take into account holidays and weeken the variation may be even greater. Thus, the production or sales for the month of February may be l not because of any real drop in activity but because of the fact that February has fewer days. Thus purpose of adjusting for calendar variation is to eliminate certain spurious differences which are cau by differences in number of days in various months. The adjustment for calendar variation is made dividing each monthly total by the number of days in the month (sometimes by the number of wor days in the month) thus arriving at daily average for each month. Comparable (adjusted) monthly may then be obtained by multiplying each of the values by 30.4167(365/12), the average number of in a month (In a leap year this factor is 30.5).

2. *Population Changes.* Certain types of data call for adjustment for population changes. Ch in the size of population can easily distort comparisons of income, production and consumption fi For example, national income may be increasing year after year, but per capita income may be dec because of greater pressure of population. Similarly, the production of a commodity may be going the per capita consumption may be declining. In such cases where it is necessary to adjust d population changes, a very simple procedure is followed, *i.e.*, the data are expressed on a per basis by dividing the original figures by the appropriate population total.

---

*Elmer C. Bratt: *Business Cycle and Forecasting.*

**J. A. Schumpeter: *Business Cycles.*

3. *Price Changes.* An adjustment for price changes is necessary whenever we have a value series and are interested in quantity changes alone. Because of rising prices the total sale proceeds may go up even when there is a fall in the number of units sold. For example, if in 2009, 1,000 units of a commodity that is priced Rs. 10 are sold, the total sale proceeds would be 1,000 × 10 = Rs. 10,000. Assume that in 2010 the price of the commodity increases from Rs. 10 to Rs. 11. If the sales do not decline, the total sale proceeds will be 1,000×11 = Rs. 11,000. This increase in sale proceeds, *i.e.*, Rs. 1,000, is not due to increase in the demand of the commodity but purely because of the rise in price from Rs. 10 to Rs. 11. Since value is equal to price per unit multiplied by the number of units sold, the effect of price changes can be eliminated by dividing each item in a value series by an approximate price index. This in fact is the process of deflating which has been discussed in the chapter on index numbers.

4. *Comparability.* For any meaningful analysis of time series, it is necessary to see that the data are strictly comparable throughout the time period under investigation. Quite often it is difficult or even impossible to get strictly comparable data. For example, if we are observing a phenomenon over the last 25 years, the comparability may be observed by differences in definition, differences in geographical average, differences in the method adopted, change in the method of reporting, etc. For example, a sale figure for January 2008 may give the average for that month, some years later the corresponding sales figure may give the total for the month or perhaps sales on the 15th or last day of the month. If such type of changes are not taken into account, the data cannot strictly be compared and its analysis would lead to fallacious conclusion.

STRAIGHT LINE TREND—METHODS OF MEASUREMENT

The following methods are used for measuring trend :

1. The Freehand or Graphic Method,
2. The Semi-average Method,
3. The Method of Least Squares, and

Each of these methods is discussed below :

1. Freehand or Graphic Method

This is the simplest method of studying trend. Under this method, the given data are plotted on a graph paper and a trend line is fitted to the data just by *inspecting the graph of the series.* There is no formal statistical criterion whereby the adequacy of such a line can be judged and it all depends on the judgment of the statistician. However, as a rough guide the line should be so drawn that it passes between the plotted points in such a manner that the fluctuations in one direction are approximately equal to those in the other direction and that it shows a general movement.

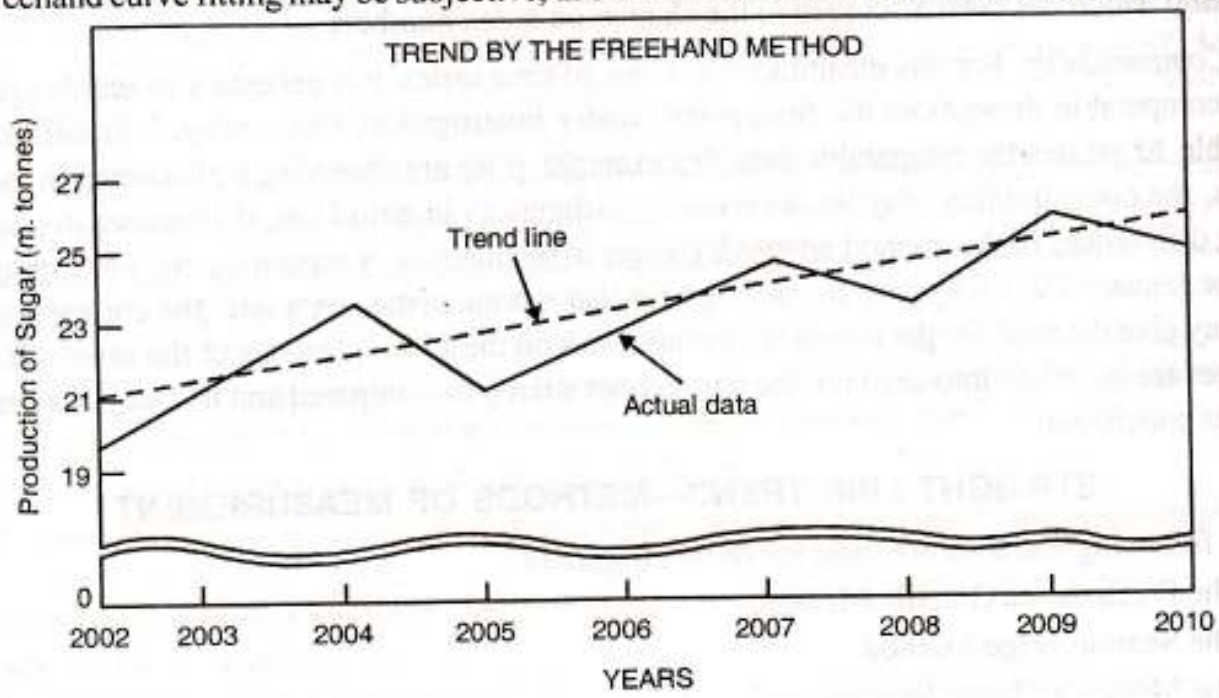
When a trend line is fitted by the freehand method, an attempt should be made to make it conform as much as possible to following conditions :

1. It should be smooth—either a straight line or a combination of long gradual curves.
2. The sum of the vertical deviations from the trend of the annual observations above the trend line should equal the sum of the vertical deviations from the trend of the observations below the trend line.
3. The sum of the squares of the vertical deviations of the observations from the trend should be as small as possible.
4. The trend should bisect the cycles so that area above the trend equals that below the trend, not only for the entire series but as much as possible for each full cycle. This last condition cannot always be met fully, but a careful attempt should be made to observe it as closely as possible.

Illustration 1. Fit a trend line to the following data by the freehand method :

Year	*Production of sugar (in million tonnes)*	*Year*	*Production of sugar (in million tonnes)*
2002	20	2007	25
2003	22	2008	23
2004	24	2009	26
2005	21	2010	25
2006	23		

The trend line drawn by the freehand method can be extended to predict future values. However, since the freehand curve fitting may be subjective, this method should not be used as a basis for prediction.



Merits and Limitations of the Freehand Method

Merits. 1. This is the simplest method of measuring trend.

2. This method is very flexible in that it can be used regardless of whether the trend is a straight line or curve.

3. The trend line drawn by a statistician experienced in computing trend and having knowledge of the economic history of the concern or the industry under analysis may be better expression of the secular movement than a trend fitted by the use of a rigid mathematical formula which, while providing a good fit to the points, may have no other logical justification. In fact a specialist of long experience who is familiar with the institutional setting, history and behaviour of the series may well be able manually to fit a trend superior to one derived by mathematical means. Although the freehand method is not recommended for beginners, it has considerable merit in the hands of experienced statisticians and is widely used in applied situations.

Limitations. 1. This method is highly subjective because the trend line depends on the personal judgment of the investigator and, therefore, different persons may draw different trend lines from the same set of data. Moreover, the work cannot be left to clerks and it must be handled by skilled and experienced people who are well conversant with the history of the particular concern.

2. Since freehand curve fitting is subjective, it cannot have much value if it is used as a basis for predictions.

3. Although this method appears simple and direct, it takes a lot of time to construct a freehand trend if a careful and conscientious job is done.

It is only after long experience in trend fitting that a statistician should attempt to fit a trend line by inspection.

2. Method of Semi-Averages

When this method is used the given data are divided into two parts, preferably, with the equal number of years. For example, if we are given data from 1993 to 2010, *i.e.*, over a period of 18 years, the two equal parts will be first nine years, *i.e.*, from 1993 to 2001 and from 2002 to 2010. In case of odd number of years like, 9, 13, 17, etc., two equal parts can be made simply by ignoring the middle year. For example, if data are given for 19 years from 1993 to 2010 the two equal parts would be from 1992 to 2000 and from 2002 to 2010—the middle year 2001 will be ignored.

After the data have been divided into two parts, an average (arithmetic mean) of each part is obtained. We thus get two points. Each point is plotted at the mid-point of the class-interval covered by the respective part and then the two points are joined by a straight line which gives us the required trend line. The line can be extended downwards or upwards to get intermediate values or to predict future values.

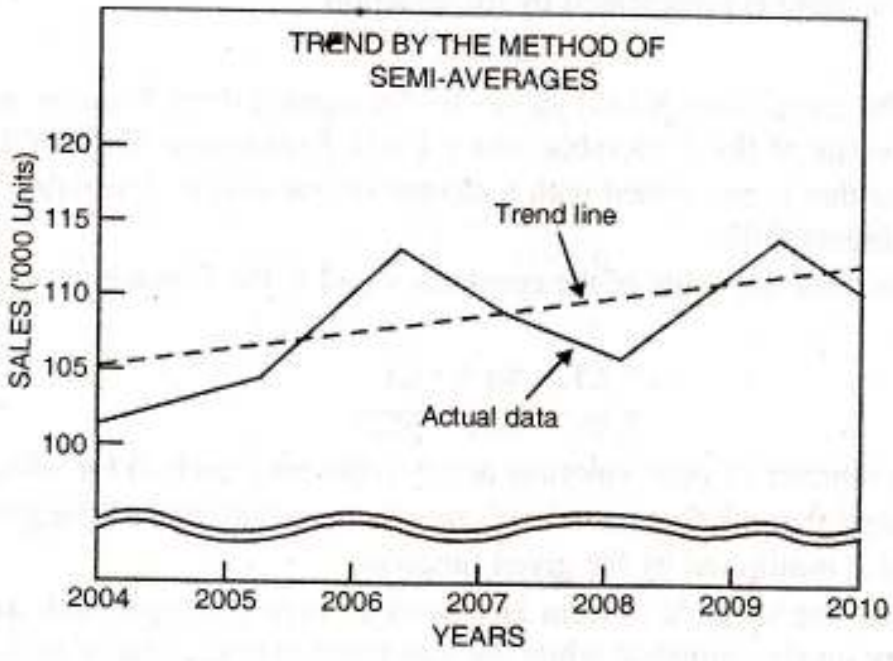
The following example shall illustrate the application of this method :

Illustration 2. Fit a trend line to the following data by the method of semi-averages :

Year	*Sales of Firm A (thousand units)*	*Year*	*Sales of Firm A (thousand units)*
2004	102	2008	108
2005	105	2009	116
2006	114	2010	112
2007	110		

Solution. Since seven years are given, the middle year shall be omitted and an average of the first three years and the last three years shall be obtained. The average of the first three years is $\frac{102+105+114}{3} = \frac{321}{3} = 107$ and the average of the last three years is $\frac{108+116+112}{3} = \frac{336}{3} = 112$. Thus, we get two points 107 and 112 which shall be plotted corresponding to their respective middle years, *i.e.*, 2005 and 2009. By joining these two points we shall obtain the required trend line. The line can be extended and can be used either for prediction or for determining intermediate values.

The actual data and the trend line are shown in the graph below.



Where there are even number of years like 6, 8, 10, etc., two equal parts can easily be formed and an average of each part obtained. However, when the average is to be centered there would be some problem in case the number of years is 8, 12, etc. For example, if the data relate to 2005-2010 which would be the middle year ? In such a case the average will be centered corresponding to 1st July, 2007, *i.e.*, middle of 2007 and 2008.

Merits and Limitations of the Semi-Average Method

Merits. 1. This method is simple to understand compared to the moving average method and method of least squares.

2. This is an objective method of measuring trend as everyone who applies the method is bound to get the same result (of course, leaving aside the arithmetical mistakes).

Limitations. 1.This method assumes straight line relationship between the plotted points regardless of the fact whether that relationship exists or not.

2. The limitations of arithmetic average shall automatically apply. If there are extremes in either half or both halves of the series, then the trend line is not a true picture of the growth factor. This danger is greatest when the time period represented by the average is small. Consequently, trend values obtained are not precise enough for the purpose either of forecasting the future trend or of eliminating trend from original data.

For the above reasons if the arithmetic average of the data is to be used in estimating the secular movement, it is sometimes better to use moving average than the semi-averages.

3. Method of Least Squares

This method is most widely used in practice. When this method is applied, a trend line is fitted to the data in such a manner that the following two conditions are satisfied :

(1) $\Sigma(Y-Y_c) = 0$

i.e., the sum of deviations of the actual values of Y and the computed values of Y is zero.

(2) $\Sigma (Y - Y_c)^2$ is least,

i.e., the sum of the squares of the deviations of the actual and computed values is least from this line. That is why this method is called the method of least squares. The line obtained by this method is known as the line of '*best fit*'.

The method of least squares can be used either to fit a straight line trend or a parabolic trend.

The straight line trend is represented by the equation

$$Y_c = a + bX$$

where Y_c denotes the trend (computed) values to distinguish them from the actual Y values, a is the Y intercept or the value of the Y variable when $X = 0$, b represents slope of the line or the amount of change in Y variable that is associated with a change of one unit in X variable. The X variable in time series analysis represents time.

In order to determine the value of the constants a and b, the following two normal equations are to be solved* :

$$\Sigma Y = Na + b\Sigma X \quad \text{...(i)}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \text{...(ii)}$$

where N represents number of years (months or any other time period) for which data are given.

It should be noted that the first equation is nearly the summation of the given function, the second is the summation of X multiplied by the given function.

We can measure the variable X from any point of time in origin such as the first year. But the calculations are very much simplified when the mid-point in time is taken as the origin because in that case the negative values in the first half of the series balance out the positive values in the second half so that $\Sigma X = 0$. In other words, the time variable is measured as a deviation from its mean. Since $\Sigma X = 0$, the above two normal equations would take the form :

*For details, see chapter on Regression Analysis.

$$\Sigma Y = Na$$
$$\Sigma XY = b\Sigma X^2$$

The values of a and b can now be determined easily.

Since $\Sigma Y = Na$,

$$a = \frac{\Sigma Y}{N} = \bar{Y}$$

Since $\Sigma XY = b\Sigma X^2$,

$$b = \frac{\Sigma XY}{\Sigma X^2}$$

The constant a give the arithmetic mean of Y and the constant b indicates the rate of change.

It should be noted that in case of odd number of years when the deviations are taken from the middle year, $\sum X$ would always be zero, provided there is no gap in the data given. However, in case of even number of years also $\sum X$ would always be zero if the X origin is placed midway between the two middle years. Hence both in odd as well as in even number of years we can use the simple procedure of determining the values of the constant a and b.

Illustration 3. Below are given the figures of production (in m. tonnes) of a sugar factory:

Year	:	2004	2005	2006	2007	2008	2009	2010
Production (in m. tonnes)	:	80	90	92	83	94	99	92

(*i*) Fit a straight line trend to these figures.

(*ii*) Plot these figures on a graph and show the trend line.

(*iii*) Estimate the likely sales of the company during 2012.

Solution : (*i*) FITTING THE STRAIGHT LINE TREND

Year	*Production (in m. tonnes)* Y	*Deviations from middle year* X	XY	X^2	*Trend Values* Y_c
2004	80	–3	–240	9	84
2005	90	–2	–180	4	86
2006	92	–1	–92	1	88
2007	83	0	0	0	90
2008	94	+1	+94	1	92
2009	99	+2	+198	4	94
2010	92	+3	+276	9	96
$N = 7$	$\Sigma Y = 630$	$\Sigma X = 0$	$\Sigma XY = 56$	$\Sigma X^2 = 28$	$\Sigma Y_c = 630$

The equation of the straight line is: $Y_c = a + bX$.

Since $\Sigma X = 0$

$$a = \frac{\Sigma Y}{N} = \frac{630}{7} = 90;\ b = \frac{\Sigma XY}{\Sigma X^2} = \frac{56}{28} = 2$$

Hence the equation of the straight line trend is :

$$Y_c = 90 + 2X.$$

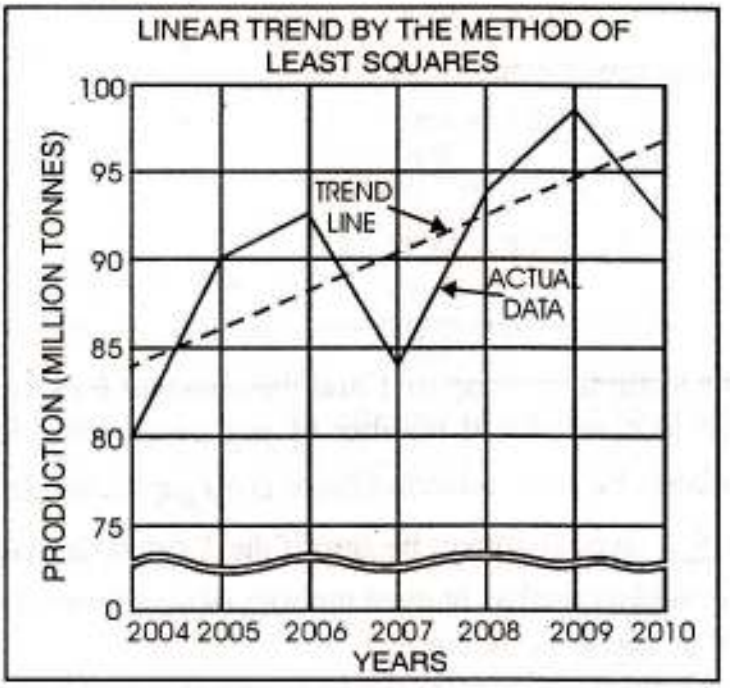
For $X = -3$, $Y_c = 90 + 2(-3) = 84$

For $X = -2$, $Y_c = 90 + 2(-2) = 86$

For $X = -1$, $Y_c = 90 + 2(-1) = 88$.

Similarly, by putting $X = 0, 1, 2, 3$ we can obtain other trend values. However, since the value of b is constant, only first trend value need be obtained and then if the value of b is positive we may continue adding the value of b to every preceding value. For example, in the above case for 2004 the calculated value of Y is 84. For 2005 it will be 84 + 2 = 86; for 2006 it will be 86 + 2 = 88, and so on. If b is negative then instead of adding we will deduct.

(*ii*) The graph of the above data is given below :



(*iii*) For 2012, X would be +5. Putting $X = 5$ in the equation

$$Y_{2012} = 90 + 2(5) = 100$$

Hence the likely production for 2012 is 100 m. tonnes.

Illustration 4. Apply the method of least squares to obtain the trend values from the following data :

Year	*Sales (in lakh tonnes)*	*Year*	*Sales (in lakh tonnes)*
2006	100	2009	140
2007	120	2010	80
2008	110		

Also predict the sales for the year 2013.

Solution. CALCULATION OF TREND VALUES BY THE METHOD OF LEAST SQUARES

Year	*Sales* Y	*Deviations from middle year* X	XY	X^2	Y_c
2006	100	–2	–200	4	114
2007	120	–1	–120	1	112
2008	110	0	0	0	110
2009	140	+1	+140	1	108
2010	80	+2	+160	4	106
$N = 5$	$\Sigma Y = 550$	$\Sigma X = 0$	$\Sigma XY = -20$	$\Sigma X^2 = 10$	$\Sigma Y_c = 550$

The equation of the straight line trend is : $Y_c = a + bX$.

Since $\Sigma X = 0$, $a = \dfrac{\Sigma Y}{N} = \dfrac{550}{5} = 110$ and $b = \dfrac{\Sigma XY}{\Sigma X^2} = \dfrac{-20}{10} = -2$

The required equation is : $Y_c = 110 - 2X$.

For $X = -2$, $Y_c = 110 - 2(-2) = 114$.

Now the other trend values will be obtained by *deducting* the value of b from the preceding value. Thus for 20 he trend value will be 114 – 2 = 112 (since the value of b is negative). For 2013, likely sales = 100 lakh tonnes (since would be 5 for 2013).

Illustration 5. Calculate the trend values by the method of least squares from the data given below and estimate sales for the year 2012-13.

Year	:	2005-06	2006-07	2007-08	2008-09	2009-10
Sales of T.V. Sets (in lakh)	:	12	18	20	23	27

Solution. CALCULATION OF TREND VALUES BY THE METHOD OF LEAST SQUARES

Year	*Sales* Y	*Taking 2007.5 as origin* X	XY	X^2	*Trend values* Y_c
2005-06	12	–2	–24	4	13.0
2006-07	18	–1	–18	1	16.5
2007-08	20	0	0	0	20.0
2008-09	23	+1	+23	1	23.5
2009-10	27	+2	+54	4	27.0
$N = 5$	$\Sigma Y = 100$	$\Sigma X = 0$	$\Sigma XY = 35$	$\Sigma X^2 = 10$	$\Sigma Y_c = 100$

The equation of the straight line trend is : $Y_c = a + bX$.

Since $\Sigma X = 0$, $a = \frac{\Sigma Y}{N} = \frac{100}{5} = 20$ and $b = \frac{\Sigma XY}{\Sigma X^2} = \frac{35}{10} = 3.5$

Thus the equation of the straight line trend is : $Y = 20 + 3.5X$

$Y_{2005\text{-}06} = 20 + 3.5\ (-2) = 13$

$Y_{2006\text{-}07} = 20 + 3.5\ (-1) = 16.5$, etc.

For 2012-13, X would be + 5.

Hence, $Y_{2012\text{-}13} = 20 + 3.5(+5) = 20 + 17.5 = 37.5$

Thus the estimated sales of television sets for the year 2012-13 is 37.5 lakh.

Illustration 6. Fit a straight line trend by the method of least squares to the following data and find the trend values :

Year	:	2005	2006	2007	2008	2009	2010
Sale of airconditioners (in lakh)	:	10	13	16	21	24	30

Solution. FITTING STRAIGHT LINE TREND BY THE METHOD OF LEAST SQUARES

Year	*Sales (in lakh)* Y	*Taking deviations from 2007* X	XY	X^2	*Trend values* Y_c
2005	10	–2	–20	4	9.143
2006	13	–1	–13	1	13.086
2007	16	0	0	0	17.029
2008	21	+1	+21	1	20.972
2009	24	+2	+48	4	24.915
2010	30	+3	+90	9	28.858
$N = 6$	$\Sigma Y = 114$	$\Sigma X = 3$	$\Sigma XY = 126$	$\Sigma X^2 = 19$	$\Sigma Y_c = 114.003$

The equation of the straight line trend is : $Y_c = a + bX$.

Since ΣX is not zero, we have to solve the two normal equations simultaneously.

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

$$114 = 6a + 3b \quad ...(i)$$

$$126 = 3a + 19b \quad ...(ii)$$

Multiplying Eqn. (*ii*) by 2,

$$114 = 6a + 3b$$

$$252 = 6a + 38b$$

$$-35b = -138 \quad \text{or } b = 3.943$$

Putting the value of b in Eqn. (*i*)

$$114 = 6a + 3\ (3.943)$$

$$6a + 11.829 = 114$$

$$6a = 114 - 11.829 = 102.171 \text{ or } a = 17.029$$

The equation of the straight line trend is :

$$Y = 17.029 + 3.943\,X$$

$$Y_{2005} = 17.029 + 3.943\,(-2) = 9.143$$

$$Y_{2006} = 9.143 + 3.943 = 13.086$$

The other trend values can be obtained similarly by adding the value of b to the preceding value.

We can simplify the above calculations in case of even number of years by taking deviations from middle, *i.e.*, 2007.5 in the above price and apply the shortcut method.

Year	*Sales (in lakhs)* Y	*Taking deviation from 2007.5*	*Multiplying deviations by 2* X	XY	X^2
2005	10	−2.5	−5	−50	25
2006	13	−1.5	−3	−39	9
2007	16	−0.5	−1	−16	1
2008	21	+0.5	+1	+21	1
2009	24	+1.5	+3	+72	9
2010	30	+2.5	+5	+150	25
$N = 6$	$\Sigma Y = 114$		$\Sigma X = 0$	$\Sigma XY = 138$	$\Sigma X^2 = 70$

Since $\Sigma X = 0$, $a = \frac{\Sigma Y}{N} = \frac{114}{6} = 19$, $b = \frac{\Sigma XY}{\Sigma X^2} = \frac{138}{70} = 1.9714$

Thus the equation of the straight line trend is : $Y_c = 19 + 1.9714X$

$Y_{2005} = 19 + 1.9714\,(-5) = 19 - 9.857 = 9.143$

$Y_{2008} = 19 + 1.9714\,(-3) = 19 - 5.9142 = 13.086$, etc.

Note. Instead of calculating like this we can double the value of b (since b is giving half-yearly trend value) and add to the preceding value.

Annual trend value of $b = 1.9714 \times 2 = 3.943$

$Y_{2006} = 9.143 + 3.943 = 13.086$

$Y_{2007} = 13.086 + 3.943 = 17.029$ (as before)

Deviations have been multiplied by 2 just to simplify the calculations.

Merits and Limitations

Merits. 1. This is a mathematical method of measuring trend and as such there is no possibility of subjectiveness.

2. The line obtained by this method is called *the line of best fit* because it is this line from where the sum of the positive and negative deviations is zero and the sum of the squares of the deviations is least, *i.e.*, $\Sigma\,(Y - Y_c) = 0$ and $\Sigma\,(Y - Y_c)^2$ least.

Limitations. Mathematical curves are useful to describe the general movement of a time series, but it is doubtful whether any analytical significance should be attached to them, except in special cases. It is seldom possible to justify on theoretical grounds any real dependence of a variable with the passage of time. Variables do change in a more or less systematic manner over time, but this can usually be attributed to the operation of other explanatory variables. Thus, many economic time series show persistent upward trends over time due to a growth of population or to a general rise in prices, *i.e.*, national income and the trend element can to a considerable extent be eliminated by expressing these series per capita or in terms of constant purchasing power. For these reasons mathematical trends are generally best regarded as tools for describing movements in time series rather than as theories of the causes of such movement. It follows that it is extremely dangerous to use trends to forecast future movements of a time series. Such forecasting, involving as it does extrapolation, can be valid only if there is theoretical justification for the particular trend as an expression of a functional relationship between the variable under consideration

and the time. But if the trend is purely descriptive of past behaviour, it can give few clues about future behaviour. Often the explanation of a trend gives ridiculous results which themselves are *prima facie* evidence that the trend could not be maintained.

Hence, mathematical methods of fitting trend are not foolproof—in fact, they can be a source of some of the most serious errors that are made in statistical work. They should never be used unless rigidly controlled by a separate logical analysis. Trend fitting depends upon the judgment of the statistician, and a skilfully made freehand sketch may often be more practical than a refined mathematical formula.*

NON-LINEAR TREND

The straight line trends discussed above indicate the increase or decrease of a time series at constant amount. It is the simplest form in describing the secular trend movement and the description of the trend is frequently accurate. However, in many cases, a straight line cannot fit the data adequately. For example, a time series may have faster (or slower) increase at early stage and have a slower (or faster increase at more recent time. In such a case better description of the time series is given by a non-linear curve rather than straight line.

The following are the methods of measuring non-linear trends:

1. Freehand or Graphic Method.
2. Moving Average Method.
3. A parabolic trend by a second degree polynomial equation obtained by the method of least squares.

1. Freehand or Graphic Method

As explained earlier, this method involves an element of subjectiveness and as such is not recommended for general use.

2. Method of Moving Averages

When a trend is to be determined by the method of moving averages, the average value for a number of years (or months, or weeks) is secured, and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the average. The effect of averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend.

While applying this method, it is necessary to select a period for moving average such as 5-yearly moving average, 6-yearly moving average, 8-yearly moving average, etc. The period of moving average is to be decided in the light of the length of the cycle. Since the moving average method is most commonly applied to data which are characterised by cyclical movements, it is necessary to select a period for moving average which coincides with the length of cycle, otherwise the cycle will not be entirely removed. This danger is more severe, the shorter the time period represented by the average. When the period of moving average and the period of the cycle do not coincide the moving average will display a cycle which has the same period as the cycle in the data, but which has less amplitude than the cycle in the data. Often we find that the cycles in the data are not of uniform length. In such a case, we should take a moving average period equal to or somewhat greater than the average period of the cycle in the data. Ordinarily the necessary period will range between three and ten years for general business series but even longer periods are required for certain types of data.

*Riggleman and Frisbee: *Business Statistics.*

The formula for 3-yearly moving average will be :

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}, \frac{d+e+f}{3}, \ldots\ldots$$

and for 5-yearly moving average

$$\frac{a+b+c+d+e}{5}, \frac{b+c+d+e+f}{5}, \frac{c+d+e+f+g}{5}, \ldots\ldots$$

Illustration 7. Calculate the 5-yearly and 7-yearly moving average for the following data of a number of commercial industrial failures in a country during 1995 to 2010.

Year	*No. of failures*	*Year*	*No. of failures*
1995	23	2003	9
1996	26	2004	13
1997	28	2005	11
1998	32	2006	14
1999	20	2007	12
2000	12	2008	9
2001	12	2009	3
2002	10	2010	1

Solution. CALCULATION OF 5-YEARLY AND 7-YEARLY MOVING AVERAGE

Year	*No. of failures*	*5-yearly moving totals*	*5-yearly moving average*	*7-yearly moving totals*	*7-yearly moving average*
1995	23	—	—	—	—
1996	26	—	—	—	—
1997	28	129	25.8 or 26	—	—
1998	32	118	23.6 = 24	153	21.9 or 22
1999	20	104	20.8 = 21	140	20.0 = 20
2000	12	86	17.2 = 17	123	17.6 = 18
2001	12	63	12.6 = 13	108	15.4 = 15
2002	10	56	11.2 = 11	87	12.4 = 12
2003	9	55	11.0 = 11	81	11.6 = 12
2004	13	57	11.4 = 11	81	11.6 = 12
2005	11	59	11.8 = 12	78	11.1 = 11
2006	14	59	11.8 = 12	71	10.1 = 10
2007	12	49	9.8 = 10	63	9.0 = 9
2008	9	39	7.8 = 8	—	—
2009	3	—	—	—	—
2010	1	—	—	—	—

If the period of moving average is even, say, four-yearly or six-yearly, the moving total and moving average which are placed at the centre of the time span from which they are computed fall between two time periods. This placement is inconvenient since the moving average so placed would not coincide with an original time period. We, therefore, synchronise moving averages and original data. This process is called centering and consists of taking a two-period moving average of the moving averages.*

* There is another method of centering the moving averages. If we are calculating 4-yearly moving average, we will first take four-yearly totals and of these totals, we will again take 2-yearly totals and divide these totals by 8.

Illustration 8. Work out the centered 4-yearly moving average for the following data :

Year	*Tonnage of cargo cleared*	*Year*	*Tonnage of cargo cleared*
1999	1102	2005	1452
2000	1250	2006	1549
2001	1180	2007	1586
2002	1340	2008	1476
2003	1212	2009	1624
2004	1317	2010	1586

Solution. CALCULATION OF THE CENTERED FOUR-YEARLY MOVING AVERAGE

Year	*Tonnage of cargo cleared*	*4-yearly moving totals*	*4-yearly moving average*	*4-yearly centered moving average*
1999	1102	—	—	—
2000	1250	—		—
	→	4872	1218.00	
2001	1180		→	1231.75
	→	4982	1245.50	
2002	1340		→	1253.87
	→	5049	1262.25	
2003	1212		→	1296.25
	→	5321	1330.25	
2004	1317		→	1356.37
	→	5530	1382.50	
2005	1452		→	1429.25
	→	5904	1476.00	
2006	1549		→	1495.87
	→	6063	1515.75	
2007	1586		→	1537.25
	→	6235	1558.75	
2008	1476		→	1563.37
	→	6272	1568.00	
2009	1624	—	—	—
2010	1586	—	—	—

Merits and Limitations

Merits. 1. This method is simple as compared to the method of least squares.

2. It is a flexible method of measuring trend. If a few more figures are added to the data, the entire calculations are not changed—we only get some more trend values.

3. If the period of moving average happens to coincide with the period of cyclical fluctuations in the data, such fluctuations are automatically eliminated.

4. The moving average has the advantage that it follows the general movements of the data and that its shape is determined by the data rather than the statistician's choice of a mathematical function.

Limitations. 1. Trend values cannot be computed for all the years. The longer the period of moving average, the greater the number of years for which trend values cannot be obtained. For example, in a three-yearly moving average, trend values cannot be obtained for the first year and last year, in five-yearly moving average for the first two years and the last two years, and so on. It is often these extreme years in which we are most interested.

2. Great care has to be exercised in selecting the period of moving average. No hard and fast rules are available for the choice of the period and one has to use his own judgment.

3. Since the moving average is not represented by a mathematical function, this method cannot be used in forecasting which is one of the main objectives of trend analysis.

4. Although theoretically we say that if the period of moving average happens to coincide with the period of cycle, the cyclical fluctuations are completely eliminated, but in practice since the cycles are by no means perfectly periodic, the lengths of the various cycles in any given series will usually vary considerably and, therefore, no moving average can completely remove the cycle. The best results would be obtained by a moving average whose period was equal to the average length of all the cycles in the given series. However, it is difficult to determine the average length of the cycle until the cycles are isolated from the series.

5. Finally, when the trend situation is not linear (a straight line) the moving average lies either above or below the true sweep of the data.

The moving average is appropriate for trend computation only when :

(*a*) the purpose of investigation does not call for current analysis or forecasting,

(*b*) the trend is linear, and

(*c*) the cyclical variations are regular both in period and amplitudes.

However, in practice, these conditions rarely hold true.

3. Second Degree Parabola

The simplest example of the non-linear trend is the *second degree parabola*, the equation of which is written in the form :

$$Y_c = a + bX + cX^2$$

When numerical values for constants a, b and c have been derived, the trend value for any year may be computed substituting in the equation the value of X for that year. The values of a, b and c can be determined by solving the following three normal equations simultaneously :

(*i*) $$\Sigma Y = Na + b\Sigma X + c\Sigma X^2$$

(*ii*) $$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

(*iii*) $$\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

Note that the first equation is merely the summation of the given function, the second is the summation of X multiplied into the given function, and the third is the summation of X^2 multiplied into the given function.

When time origin is taken between two middle years ΣX and ΣX^3 would be zero. In that case the above equations are reduced to :

(*i*) $$\Sigma Y = Na + c\Sigma X^2$$

(*ii*) $$\Sigma XY = b\Sigma X^2$$

(*iii*) $$\Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4$$

The value of b can now directly be obtained from equation (*ii*) and that of a and c by solving (*i*) and (*iii*) simultaneously. Thus,

$$a = \frac{\Sigma Y - c\Sigma X^2}{N} \; ; \; b = \frac{\Sigma XY}{\Sigma X^2}$$

$$c = \frac{N\Sigma X^2 Y - \Sigma X^2 \Sigma Y}{N\Sigma X^4 - (\Sigma X^2)^2}$$

Illustration 9. The price (in Rs.) of a commodity during 2005-2010 is given below. Fit a parabola $Y = a+bX+cX^2$ to this data. Estimate the price of commodity for the year 2013.

Year	*Price*	*Year*	*Price*
2005	100	2008	140
2006	107	2009	181
2007	128	2010	192

Also plot the actual and trend values on the graph.

Solution. To determine the value of a, b and c, we solve the following normal equations :

$$\Sigma Y = Na + b\Sigma X + c\Sigma X^2$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$
$$\Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

Year	*Price* Y	X	X^2	X^3	X^4	XY	X^2Y	*Trend Values* Y_c
2005	100	–2	4	–8	16	–200	400	97.744
2006	107	–1	1	–1	1	–107	107	110.426
2007	128	0	0	0	0	0	0	126.680
2008	140	+1	1	+1	1	+140	140	146.506
2009	181	+2	4	+8	16	+362	724	169.904
2010	192	+3	9	+27	81	+576	1728	196.874
$N = 6$	$\Sigma Y = 848$	$\Sigma X = 3$	$\Sigma X^2 = 19$	$\Sigma X^3 = 27$	$\Sigma X^4 = 115$	$\Sigma XY = 771$	$\Sigma X^2Y = 3099$	$\Sigma Y_c = 848.134$

$$848 = 6a + 3b + 19c \quad ...(i)$$
$$771 = 3a + 19b + 27c \quad ...(ii)$$
$$3{,}099 = 19a + 27b + 115c \quad ...(iii)$$

Solving Eqns. (*i*) and (*ii*), we get

$$35b + 35c = 694 \quad ...(iv)$$

Multiplying Eqn. (*ii*) by 19 and Eqn. (*iii*) by 3 and subtracting, we get

$$53.52 = 280b + 168c \quad ...(v)$$

Solving Eqns. (*iv*) and (*v*), we get

$$c = 1.786$$

Substituting the value of c in Eqn. (*iv*), we get

$$b = 18.04$$

Putting the value of b and c in Eqn. (*i*), we get

$$a = 126.68$$

Thus, $a = 126.68$, $b = 18.04$ and $c = 1.786$

Substituting the values in the equation

$$Y = 126.68 + 18.04X + 1.786X^2$$

When $X = -2$, $Y = 126.68 + 18.04(-2) + 1.786(-2)^2$
$= 126.68 - 36.08 + 7.144 = 97.744$

When $X = -1$, $Y = 126.68 + 18.04(-1) + 1.786(-1)^2$
$= 126.68 - 18.04 + 1.786 = 110.426$

When $X = 0$, $Y = 126.68$

When $X = 1$, $Y = 126.68 + 18.04 + 1.786 = 146.506$

When $X = 2$, $Y = 126.68 + 18.04(2) + 1.786(2)^2$
$Y = 126.68 + 36.08 + 7.144 = 169.904$

When $X = 3$, $Y = 126.68 + 18.04(3) + 1.786(3)^2$
$= 126.68 + 54.12 + 16.074 = 196.874$

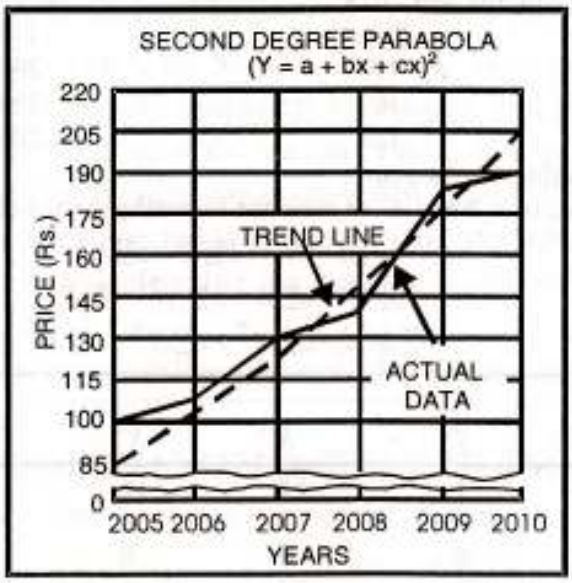
Price for 2013

For 2013 X will be 6.

When $X = 6$ $Y = 126.68 + 18.04(6) + 1.786(6)^2$
$= 126.68 + 108.24 + 64.296 = 299.216$

Thus, the likely price of the commodity for the year : 2013 is Rs. 299.216.

The graph of the actual trend values is given below :



I. MEASURING TRENDS BY LOGARITHMS

The trends discussed so far were plotted on arithmetic scales. Trends may also be plotted on a semi-logarithmic (or semi-log) chart in the form of a straight line or a nonlinear curve. A straight line on the semi-log chart shows the increase of Y values of a time series at a constant rate (A straight line on an arithmetic chart indicates the increase at a constant amount). When it is a nonlinear curve on the semi-log chart an upward curve shows the increase at varying rates, depending on the shape of the slope—the steeper the slope, the higher is the rate of increase.

The types of trend usually computed by logarithms are :

1. Exponential trends, and
2. Growth curves.

Exponential Trends

The equation of the exponential curve is of the form

$$Y = ab^X$$

Putting the equation in logarithmic form, we get

$$\log Y = \log a + X \log b$$

When plotted on a semi-logarithmic graph, the curve gives a straight line. However, on an arithmetic chart the curve gives a nonlinear trend. In order to find out the values of a and b, the two normal equations to be solved are :

$$\Sigma \log Y = N \log a + \log b \, \Sigma X$$

$$\Sigma (X.\log Y) = \log a \, \Sigma X + \log b \, \Sigma X^2$$

When deviations are taken from middle year, *i.e.*, $\Sigma X = 0$, the above equations take the following form

$$\Sigma \log Y = N \log a$$

$$\Sigma (X.\log Y) = \log b \, \Sigma X^2$$

or

$$\log a = \frac{\Sigma \log Y}{N}; \text{ and } \log b = \frac{\Sigma (X . \log Y)}{\Sigma X^2}$$

Take the antilogs of these expressions to arrive at the actual trend values.

Illustration 10. The sales of a company in lakhs of rupees for the year 2004 to 2010 are given below :

Years :	2004	2005	2006	2007	2008	2009	2010
Sales : (m. tonnes)	32	47	65	92	132	190	275

Estimate sales figures for the year 2013 using an equation of the form $Y = ab^X$, where X = years and Y = sales.

Solution. FITTING EQUATION OF THE FORM $Y = ab^X$

Year	*Sales* Y	*Deviations from 2007* X	X^2	*log Y*	*X log Y*
2004	32	–3	9	1.5051	–4.5153
2005	47	–2	4	1.6721	–3.3442
2006	65	–1	1	1.8129	–1.8129
2007	92	0	0	1.9638	0
2008	132	+1	1	2.1206	+2.1206
2009	190	+2	4	2.2788	+4.5576
2010	275	+3	9	2.4393	+7.3179
$N = 7$	$\Sigma Y = 833$	$\Sigma X = 0$	$\Sigma X^2 = 28$	$\Sigma \log Y = 13.7926$	$\Sigma X \log Y = 4.3237$

We have to fit the equation $Y = ab^X$. It can be written as

$$\log Y = \log a + X \log b$$

Since deviations are taken from middle year, $\Sigma X = 0$

$$\log a = \frac{\Sigma \log Y}{N} = \frac{13.7926}{7} = 1.97$$

$$\log b = \frac{\Sigma (X \log Y)}{\Sigma X^2} = \frac{4.3237}{28} = 0.1544$$

Hence $\log Y = 1.97 + 0.1544\, X$

For 2013, X would be + 6.

$$\log Y = 1.97 + 0.1544\,(6) = 2.8964$$

$$Y = \text{Antilog } 2.8964 = 787.77$$

Thus the estimated figure of sales for the year 2013 is 787.77 m. tonnes.

Illustration 11. Fit a logarithmic straight line to the following data :

Years :	2005	2006	2007	2008	2009	2010
Production (m. tonnes of steel) :	64	70	75	82	88	95

Solution. FITTING OF LOGARITHMIC STRAIGHT LINE

Year	*Production* Y	X	*log Y*	X^2	*X.log Y*
2005	64	–3	1.8062	9	–5.4186
2006	70	–2	1.8451	4	–3.6902
2007	75	–1	1.8751	1	–1.8751
2008	82	0	1.9138	0	0
2009	88	+1	1.9445	1	+1.9445
2010	95	+2	1.9777	4	+3.9554
$N = 6$	$\Sigma Y = 474$	$\Sigma X = -3$	$\Sigma \log Y = 11.3624$	$\Sigma X^2 = 19$	$\Sigma X.\log Y = -5.084$

The logarithmic straight line trend is given by

$$\log Y = \log a + X \log b$$

The two normal equations are :

$$\Sigma \log Y = N \log a + \log b\, \Sigma X$$

$$\Sigma X \log Y = \log a\, \Sigma X + \log b\, \Sigma X^2$$

Substituting the values

$$11.3624 = 6 \log a - 3 \log b \quad \ldots(i)$$

$$-5.084 = -3 \log a + 19 \log b \quad \ldots(ii)$$

Multiplying eq. (*ii*) by 2 and adding to (*i*),

$$\begin{aligned} 11.3624 &= 6 \log a - 3 \log b \\ -10.168 &= -6 \log a + 38 \log b \\ \hline 35 \log b &= 1.1944 \end{aligned}$$

$$\log b = \frac{1.1944}{35} = 0.034$$

Putting the value of $\log b$ in eqn. (*i*)

$$11.3624 = 6 \log a - 0.102$$
$$6 \log a = 11.4644$$
$$\log a = 1.911$$

Hence $$\log Y = 1.911 - 0.034 X$$

Second Degree Curves Fitted to Logarithms

We may come across data which when plotted on semi-logarithmic graph paper may continue to show curvature, being concave either upward or downward; or in other words, the ratio of change may be either increasing or decreasing. In such cases, we may fit second degree curve to the logarithms of the Y values using

$$\log Y = \log a + X \log b + X^2 \log c$$

Taking the X origin at the middle of the period, the three normal equations are:

(*i*) $$\Sigma \log Y = N \log a + \log c \Sigma X^2$$

(*ii*) $$\Sigma (X. \log Y) = \log b \Sigma X^2$$

(*iii*) $$\Sigma (X^2. \log Y) = \log a \Sigma X^2 + \log c \Sigma X^4$$

Growth Curves

In economic data very often we come across phenomenon where at first the growth is very slow, but as the product is accepted the demand increases by a greater amount each year and finally as the market becomes more and more fully developed, the amount of growth each year becomes less. The curve continues to grow more and more slowly, approaching an upper limit but not reaching it. Such series are best represented by growth curves. The growth curves do not reach a maximum and turn down in the manner of the second degree parabola.

A number of different growth curves have been used to measure secular trend, but the curves used most widely to describe growth are the *Gompertz Curve* and the *Peart Reed* or *logistic* curve.

The equation of the Gompertz curve is

$$Y = kab^X$$

which when put to logarithmic form becomes

$$\log Y = \log k + (\log a) b^X$$

The Gompertz curve serves to describe the series which while increasing seem to approach some maximum value as a limit. Although the growth continues it does so at a decreasing rate.

The equation of the logistic (or the Peart-Reed) curve is:

$$Y = \frac{1}{k + ab^X}$$

where k, a and b are constants. The logistic curve has been applied widely to population data of various kinds, both human and non-human, and it has also provided a good fit to many economic series pertaining to industrial growth.

Both the Gompertz and the logistic curves approach a finite limit. This fact ought to be taken into account when fitting a given time series to one of the curves. Often a key resource is known to exist to some finite amount, and this can be used to establish a limit on the growth of the time series.

question. Increasingly, for example, new cities are being planned with an eye towards limiting growth with planned land available having an upper limit. Whenever a given time series increasing at a constant rate but is understood to be approaching a finite limit in a predictable manner, growth curve may be appropriate for assessing the secular trend component of the series.

Conversion of Annual Trend Values to Monthly Trend Values

Usually for trend computations annual figures are employed. However, it is sometimes required to obtain monthly trend ordinates. In converting straight line trends from an annual to a monthly basis, two situations must be clearly distinguished. For series such as sales, production or earnings, the annual figure is the total of monthly figures. Here it is necessary to divide both a and b by 12 to reduce them to monthly level. In other words, on the average, monthly sales or production is one-twelth of the annual total. The b values must then be divided by 12 once again in order to convert from annual to monthly increments.

The necessity of dividing b twice by 12, that is by 144 altogether, must be clearly understood. The division of annual change by 12 gives us only the change from same month in a given year to the corresponding month in the following year, or the annual change in monthly magnitudes. However, we are here seeking for an expression of the change in each and every month, that is, monthly change in monthly magnitude. Thus, b must be divided again by 12.

In conclusion, to convert an annual trend equation to a monthly basis when the original data are given as totals, a is divided by 12 and b is divided by 144.

If X in the trend line equation represents only 6 months, it is divided by 72 instead of 144.

Illustration 12. Convert the following annual trend equation for tea production in India to a monthly trend equation :

$$Y = 108 + 1.58X$$

(Origin 2010, time unit one year, Y = tea production in million kg.)

Solution. Monthly trend equation will be obtained by dividing a by 12 and b by 144. Thus the monthly trend equation will be

$$Y = \frac{108}{12} + \frac{1.58}{144}$$

(Origin July 1, 2010, time unit 1 month, Y monthly production in million kg.)

Where data are given as monthly average per year, the value of the constant 'a' in the annual trend equation is the arithmetic mean of the twelve month total. In other words, it is already at the monthly level. The value of 'b' now represents the annual change in month magnitude. As a result, to convert an annual trend equation when annual data are expressed as monthly averages, a would remain unchanged and b is divided by 12 only.

Shifting the Trend Origin

In computing trends, the middle of the time series is often used as the origin in order to cut-short the computations. But very often we need to change the origin of the trend equation to some other point in the series. This is either to facilitate comparison of trend values among neighbouring years or to convert a trend equation from any annual to a monthly basis. Shifting the origin is a very simple matter. For example, consider the trend equation

$$Y = 110 - 2X$$

(Origin 2005, time unit 1 year)

If we wish to shift the trend equation to 2010, we note that this year precedes the stated origin of 2005 by 7 time units. Consequently, we must deduct 7 times annual increment that is $b(-7)$, from the trend value of 2005 as below :

$$Y = 110 - 2(-7) = 110 + 14 = 124.$$

The value 124 becomes the trend value at the new origin 2010 and the trend equation may now be written as

$$Y = 124 - 2X.$$

Selecting Type of Trend

We have discussed different ways of fitting trends. However, it is not all—some other equations might also be reasonably used. Even though each series presents its own individual problem, most series can be handled by the methods which we have described. Of course, what we try to do in any particular case is to select that equation or that method of measuring trend which best describes the gradual and consistent pattern of growth.

The choice of a particular type of equation that best describes the data is often difficult and needs considerable amount of judgment and experience.

While deciding the type of trend, the first step consists of plotting the data on arithmetic paper. If the trend is not linear but either:

(*a*) concave upwards,or

(*b*) concave downwards

the data should be plotted on a semi-logarithmic paper. Examination of the plotted data often provides an adequate basis for deciding upon the type of trend to use. When further guidance is needed an approximate trend may be drawn by inspection and the following tests applied to the smoothed curve*.

1. If the first differences tend to be constant, use a straight line.
2. If the second differences tend to be constant, use a second degree curve.
3. If the approximate trend when plotted on arithmetic paper is a straight line, use a straight line.
4. If the first differences of the logarithms are constant, use an exponential curve.
5. If the second differences of the logarithms are constant, fit a second degree curve to the logarithms.
6. If the first differences tend to decrease by a constant percentage, use a modified exponential.
7. If the first differences resemble a skewed frequency curve, use a Gompertz curve or a more complex logistic curve.

Choice of the Trend Period

In order to simplify the discussion of trend computation, the illustrations in this chapter are based on 7 or 8 years data only. However, wherever possible, the period should be longer. The longer the period, the less the trend values will be distorted by cyclical or random influence. The period employed should encompass a number of business cycles and should begin and end in such a way that distortion is avoided. The purpose can be accomplished by using a period that starts and finishes either in prosperity or depression, or by beginning during recovery and ending during recession.

Trend Extrapolation

Trend analysis is often employed to forecast future levels of time series data. By substituting in the trend equation values of X for the dates for which forecast are desired, the ordinates for those dates can be obtained. This process is called extrapolation. Utmost care must be exercised while interpreting such forecasts. The following points are worth considering :

(1) The forecast obtained through trend analysis is a forecast of trend only. Actual values may be expected to diverge from it because of cyclical and random factors.

---

*Croxton and Cowden: *Applied General Statistics.*

(2) The forecast has meaning only if the same basic influence that shaped the trend in the past can be expected to be controlling in the future. Thus extrapolation implicitly assumes that the institutional pattern will be quite stable. However, in practice, the extrapolation of trends goes far wide off the mark because significant changes in conditions governing the values of the data cannot be or are not properly anticipated.

(3) The type of curve employed must not only properly describe the past movement of the data but also be capable of sensible extrapolation. The extrapolation of some trends especially the more complex curves may yield results that are essentially meaningless.

II. MEASUREMENT OF SEASONAL VARIATIONS

Most of the phenomena in economics and business show seasonal patterns. When data are expressed annually, there is no seasonal variation. However, monthly or quarterly data frequently exhibit strong seasonal movements and considerable interest attaches to devising a pattern of *average seasonal variation.* For example, if we observe the sales of a bookseller, we find that of the quarter July-September (when most of the students purchase books), sales are maximum. If we know by how much the sales of this quarter are usually above or below the previous quarter for seasonal reasons, we shall be able to answer a very basic question, namely, was this due to an underlying upward tendency or simply because this quarter is usually seasonally higher than the previous quarter ?

In order to analyse seasonal variation, it is necessary to assume that the seasonal pattern is superimposed on a series of values and independent of these in the sense that the same pattern is superimposed irrespective of the level of the series, *i.e.,* the June quarter always contributes so much more or so much less of the series.

Before attempting to measure seasonal variation certain preliminary decisions must be made. For example, it is necessary to decide whether weekly, quarterly or monthly indexes are required. This will be decided in the light of the nature of the problem and the type of data available.

To obtain a statistical description of a pattern of seasonal variation it will be desirable to first free the data from the effect of trend, cycles and irregular variation. Once these other components have been eliminated we can calculate, in index form, a measure of seasonal variations which is usually referred to as seasonal index. Thus the measures of seasonal variation are called *seasonal indexes* (per cent).

For monthly data, a seasonal index consists of 12 numbers, one for each month of a year, or number of years, that has taken place typically in each month. Thus a second index may be specific or typical. A *specific seasonal index* refers to the seasonal changes during a particular year. A typical seasonal index is obtained by averaging a number of specific seasonals. It is thus a generalised expression of seasonal variations for a series. Seasonal indexes are given as percentages of their average, *i.e.,* each month is represented by a figure expressing it as a percentage of the average month. For example, if a seasonal index for January is 75, this means that for the month of January, sales, orders, purchases or whatever our data happen to be are 75 per cent of those of the average month.

There are several methods of measuring seasonal variation. However, the following methods that are more popularly used in practice are discussed below :

1. Method of Simple Averages (Weekly, Monthly or Quarterly),
2. Ratio-to-Trend Method,
3. Ratio-to-moving Average Method,
4. Link Relatives Method.

1. Method of Simple Averages

This is the simplest method of obtaining a seasonal index. The following steps are necessary in calculating the index :

(*i*) Average the unadjusted data by years and months (or quarters if quarterly data are given).

(*ii*) Find totals of January, February, etc.

(*iii*) Divide each total by the number of years for which data are given. For example, if we are given monthly data for five years then we shall first obtain total for each month for five years and divide each total by 5 to obtain an average.

(*iv*) Obtain an average of monthly averages by dividing the total of monthly averages by 12.

(*v*) Taking the average of monthly averages as 100, compute the percentage of various monthly averages as follows :

Seasonal Index for January

$$= \frac{\text{Monthly average for January}}{\text{Average of monthly averages}} \times 100$$

If, instead of the average of each month, the totals of each month are obtained, we will get the same result.

The following example shall illustrate the method.

Illustration 13. Consumption of monthly electric power in million of Kw hours for street lighting in one of the states in India during 2006-2010 is given below :

Year	*Jan.*	*Feb.*	*Mar.*	*Apr.*	*May*	*June*	*July*	*Aug.*	*Sept.*	*Oct.*	*Nov.*	*Dec.*
2006	318	281	278	250	231	216	223	245	269	302	325	347
2007	342	309	299	268	249	236	242	262	288	321	342	364
2008	367	328	320	287	269	251	259	284	309	345	367	394
2009	392	349	342	311	290	273	282	305	328	364	389	417
2010	420	378	370	334	314	296	305	330	356	396	422	452

Find out seasonal variation by the method of monthly averages.

Solution. COMPUTATION OF SEASONAL INDICES BY THE METHOD OF MONTHLY AVERAGES

Month	*Consumption of monthly electric power*					*Monthly total for 5 years*	*Five-yearly average*	*Percentage*
	2006	*2007*	*2008*	*2009*	*2010*			
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*
Jan.	318	342	367	392	420	1,839	367.8	116.1
Feb.	281	309	328	349	378	1,645	329.0	103.9
March	278	299	320	342	370	1,609	321.8	101.6
April	250	268	287	311	334	1,450	290.0	91.6
May	231	249	269	290	314	1,353	270.6	85.4
June	216	236	251	273	296	1,272	254.4	80.3
July	223	242	259	282	305	1,311	262.2	82.8
Aug.	245	262	284	305	330	1,426	285.2	90.1
Sept.	269	288	309	328	356	1,550	310.0	97.9
Oct.	302	321	345	364	396	1,728	345.6	109.1
Nov.	325	342	367	389	422	1,845	369.0	116.5
Dec.	347	364	394	417	452	1,974	394.8	124.7
					Total	19,002	3,800.4	1,200
					Average	1.5835	316.7	100

The above calculations are explained below :

1. Column No. 7 gives the total for each month for five years.
2. In column No. 8 each total of column No. 7 has been divided by 5 to obtain an average for each month.
3. The average of monthly averages is obtained by dividing the total of monthly averages by 12.

4. In column No. 9 each monthly average has been expressed as a percentage of the average of monthly averages. Thus, the percentage for January

$$= \frac{367.8}{316.7} \times 100 = 116.1$$

$$\text{Percentage for February} = \frac{329.0}{316.7} \times 100 = 103.9$$

If instead of monthly data, we are given weekly or quarterly data, we shall compute weekly or quarterly averages by following the same procedure as explained above.

Merits and Limitations of the Method of Monthly Averages

This method is the simplest of all methods of measuring seasonality. However, it is not a very good method. It assumes that there is no trend component in the series, *i.e.*, $O = CSI$. But this is not a justified assumption. Most economic series have trends and, therefore, the seasonal index computed by this method is actually an index of trends and seasonals. Furthermore, the effects of cycles on the original values may or may not be eliminated by the averaging process. This depends on the duration of the cycle and the term of the average, that is, on the number of months included in the average. Thus, this method is seldom of any value. In its simplest form, the method only serves the purpose where no definite trend exists.

2. Ratio-to-Trend Method

This method of calculating a seasonal index (also known as the percentage-to-trend method) is relatively simple and yet an improvement over the method of simple average explained in the preceding section. This method assumes that seasonal variation for a given month is constant fraction of trend. The ratio-to-trend method presumably isolates the seasonal factors in the following manner. Trend is eliminated when the ratios are computed. In effect :

$$\frac{T \times S \times C \times I}{T} = S \times C \times I$$

Random elements are supposed to disappear when the ratios are averaged. A careful selection of the period of years used in the computation is expected to cause the influences of prosperity or depression to offset each other and thus remove the cycle. For series that are not subject to pronounced cyclical or random influences and for which trend can be computed accurately, this method may suffice. The steps in the computation of seasonal index by this method are :

1. Trend values are obtained by applying the method of least squares.

2. The next step is to divide the original data month by month by the corresponding trend values and to multiply these ratios by 100. The values so obtained are now free from trend and the problem that remains is to free them also of irregular and cyclical movements.

3. In order to free the values from irregular and cyclical movements, the figures given for the various years for January, February, etc., are averaged with any one of the usual measures of central value, for instance, the *median* or the *mean*. If the data are examined month by month, it is sometimes possible to ascribe a definite cause to usually high or low values. When such causes are found to be associated with irregular variations (extremely bad weather, an earthquake, famine and the like) they may be cast out and the mean of the remaining items is referred to as a *modified mean*. Since such scrutiny of the data requires considerable knowledge of prevailing condition and is to a large extent subjective, it is often described to use the *median* which is generally not affected by very high or very low values.

4. The seasonal index for each month is expressed as a percentage of the average month. The sum of 12 values must equal 1,200 or 100%. If it does not, an adjustment is made by multiplying each index by a suitable factor $\left(\frac{1200}{\text{the sum of 12 values}}\right)$. This gives the final seasonal index.

Illustration 14. Find seasonal variations by ratio-to-trend method from the data given below :

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2006	30	40	36	34
2007	34	52	50	44
2008	40	58	54	48
2009	54	76	68	62
2010	80	92	86	82

Solution. For determining seasonal variation by ratio-to-trend method, first we will determine the trend of yearly data then convert it to quarterly data.

CALCULATING TREND BY METHOD OF LEAST SQUARES

Year	*Yearly totals*	*Yearly average* Y	*Deviations From mid-year* X	XY	X^2	*Trend val* Y_c
2006	140	35	–2	–70	4	32
2007	180	45	–1	–45	1	44
2008	200	50	0	0	0	56
2009	260	65	+1	+65	1	68
2010	340	85	+2	+170	4	80
		$\Sigma Y = 280$	$\Sigma X = 0$	$\Sigma XY = 120$	$\Sigma X^2 = 10$	

The equation of the straight line trend is $Y = a + bX$.

Since $\Sigma X = 0$, $a = \frac{\Sigma Y}{N} = \frac{280}{5} = 56$; $\frac{\Sigma XY}{\Sigma X^2} = \frac{120}{10} = 12$

Quarterly increment $= \frac{12}{4} = 3$

Calculation of Quarterly Trend Values. Consider 2006. Trend value for the middle *i.e.*, half of 2nd and half of 3rd is 32. Quarterly increment is 3. So the trend value of 2nd q 32–3/2, *i.e.*, 30.5 and for 3rd quarter is 32 + 3/2, *i.e.*, 33.5. Trend value for the 1st quart – 3, *i.e.*, 27.5 and of 4th quarter is 33.5 + 3, *i.e.*, 36.5. We thus get quarterly trend va given values are to be expressed as the percentages of the corresponding trend values.

TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Qua*
2006	27.5	30.5	33.5	36.5
2007	39.5	42.5	45.5	48.5
2008	51.5	54.5	57.5	60.
2009	63.5	66.5	69.5	72.
2010	75.5	78.5	81.5	84

QUARTERLY VALUES AS % OF TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Q*
2006	109.1	131.1	107.5	9
2007	86.1	122.4	109.9	9
2008	77.7	106.4	93.9	
2009	85.0	114.3	97.8	
2010	106.0	117.1	105.5	
Total	463.9	591.3	514.6	4
Average	92.78	118.26	102.92	
Seasonal Index adjusted	92.0	117.4	102.1	

Total of averages = 92.78+118.26 + 102.92 + 89.12 = 403.08

Seasonal Index adjusted for Ist quarter = $\frac{92.78}{100.77}$ (avg. of total averages)

Since the total is more than 400 an adjustment is made by multiplying each average by $\frac{400}{403.08}$ and final indices are obtained.

Merits and Limitations of the Ratio-to-Trend Method

Merits. (1) Compared with the method of monthly averages this method is certainly a more logical procedure for measuring seasonal variations. It has an advantage over the moving average procedure too, for it has a ratio-to-trend value for each month for which data are available. Thus there is no loss of data as occurs in the case of moving averages. This is a distinct advantage, especially when the period covered by time series is very short.

(2) It is simple to compute and easy to understand.

Limitations. The main defect of the ratio-to-trend method is that if there are pronounced cyclical swings in the series, the trend—whether a straight line or a curve—can never follow the actual data as closely as a 12-month moving average does. In consequence a seasonal index computed by the ratio-to-moving average method may be less biased than the one calculated by the ratio-to-trend method.

3. Ratio-to-Moving Average Method*

The ratio-to-moving average, also known as the percentage of moving average method, is the *most widely* used method of measuring seasonal variations. The steps necessary for determining seasonal pattern by this method are :

1. Eliminate seasonality from the data by ironing it out from the original data. Since seasonal variations recur every year—that is, since the fluctuations have a time span of 12 months—a centered 12-month moving average tends to eliminate these fluctuations. (In case of quarterly data, a centered 4 quarter moving average must be used.) The centered 12-month moving average which aims to eliminate seasonal and irregular fluctuations (S and I) represents the remaining elements of the original data, namely, trend and cycles. Thus, the centered 12-month moving average approximates $T.C$.

2. Express the original data for each month as percentage of the centered 12-month moving average corresponding to it.

3. Divide each monthly item of the original data by the corresponding 12-month moving average, and list the quotients as 'Percent of Moving Average'. We have now succeeded in eliminating from the original data to a considerable extent the disturbing influence of trend and cycles. It remains to rid the data of irregular variations. By averaging these percentages, for a given month (step 4) the irregular factors tend to cancel out and the average itself reflects the seasonal influence alone.

4. The purpose of this step is to average, and—in process of averaging—to eliminate the irregular factors. We assume that the relatively high or extremely low values of seasonal relatives for any month are caused by irregular factors. The elimination of extremes may be achieved while we are averaging all Januarys, Februarys and the like. We do this by using an appropriate type to average. The median is appropriate since it is not affected by extremes. Thus, by using the median as an average we can obtain the typical seasonal relative for each month which will not be affected by irregular factors.

Sometimes a so-called modified mean is used as an average for each month. Here, extreme values are omitted before the arithmetic mean is taken. In an array of seasonal relatives for each month, a value

*The computation by this method is identical with computations of the ratio-to-trend seasonal index just described, except that a moving average trend is substituted for the least square trend used in the previous calculation.

or several values on one end or both ends may be dropped and then the arithmetic mean of the remaining seasonal relatives is taken. A separate table is prepared in which the calculations involved in this step are shown. These means are preliminary seasonal indexes. They should average 100 per cent or total 1,200 for 12 months by definition.

5. If the total is not equal to 1,200 or 100 per cent, an adjustment is made to eliminate the discrepancy. The adjustment consists of multiplying average of each month obtained in step 4 by

$$\frac{1{,}200}{\text{The total of the modified mean for 12 months}}$$

This adjustment is made not only to achieve accuracy, but also because when we come to eliminate seasonality from the original data we do not wish to raise or lower the level of the data unduly. Thus, if a seasonal index aggregates more than 1,200 (or averages more than 100) then the original data adjusted in terms of it will total less than the unadjusted original data. If it totals less than 1,200, the opposite would be true.

The logical reasoning behind this method follows from the fact that 12-month moving average can be considered to represent the influence of cycle and trend $C \times T$. If the actual value for any month is divided by the 12-month moving average centred to that month, presumably cycle and trend are removed. This may be represented by the following expression :

$$\frac{T \times S \times C \times I}{T \times C} = S \times I$$

Thus the ratio to the moving average, from which this method gets its name, represents irregular and seasonal influences. If the ratios for each worked over a period of years are then averaged most random influences will usually be eliminated. Hence, in effect,

$$\frac{S \times I}{I} = S$$

Illustration 15. Apply ratio to moving average method to ascertain seasonal indices from the following data :

Year and month	*Sales (in thousand units)*	*Year and month*	*Sales (in thousand units)*
2007		2009	
Jan.	10	Jan.	10
Feb.	12	Feb.	12
March	13	March	11
April	15	April	12
May	16	May	13
June	16	June	15
July	17	July	15
Aug.	18	Aug.	17
Sept.	18	Sept.	18
Oct.	19	Oct.	20
Nov.	22	Nov.	22
Dec.	22	Dec.	24
2008		2010	
Jan.	11	Jan.	12
Feb.	11	Feb.	13
March	12	March	13
April	13	April	15
May	14	May	16
June	14	June	18
July	15	July	20
Aug.	15	Aug.	20
Sept.	15	Sept.	21
Oct.	16	Oct.	22
Nov.	18	Nov.	24
Dec.	20	Dec.	25

Solution. COMPUTATION OF 12-MONTH MOVING AVERAGE

Year and month	*Sales (Thousand units)*	*12-Month moving total*	*12-Month moving average*	*2-Month moving total of col.4*	*Centered 12-month moving average (col. 5 ÷ 2)*	*Percentage of centered 12-month moving average (col. 2 ÷ col. 6)*
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
2007						
Jan.	10					
Feb.	12					
March	13					
April	15					
May	16					
June	16					
		198	16.50			
July	17			33.03	16.54	102.8
		199	16.58			
Aug.	18			33.03	16.54	108.8
		198	16.50			
Sept.	18			32.92	16.46	109.4
		197	16.42			
Oct.	19			32.67	16.33	116.3
		195	16.25			
Nov.	22			32.33	16.16	136.1
		193	16.08			
Dec.	22			32.00	16.00	137.5
		191	15.92			
2008						
Jan.	11			31.67	15.83	69.5
		189	15.75			
Feb.	11			31.25	15.62	70.4
		186	15.50			
March	12			30.75	15.37	78.1
		183	15.25			
April	13			30.25	15.12	86.0
		180	15.00			
May	14			29.67	14.83	94.4
		176	14.67			
June	14			29.17	14.59	96.0
		174	14.50			
July	15			28.82	14.46	103.7
		173	14.42			
Aug.	15			28.92	14.46	103.7
		174	14.50			
Sept.	15			28.92	14.46	103.7
		173	14.42			
Oct.	16			28.75	14.37	111.3
		172	14.33			
Nov.	18			28.58	14.29	126.0
		171	14.25			
Dec.	20			28.58	14.29	140.00

Year and month	Sales (Thousand units)	12-Month moving total	12-Month moving average	2-Month moving total of col.4	Centered 12-month moving average (col. 5 ÷ 2)	Percentage of centered 12-month moving average (col. 2 ÷ col. 6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
2009						
Jan.	10	172	14.33			
				28.66	14.33	70.0
		172	14.33			
Feb.	12			28.83	14.41	83.3
		174	14.50			
March	11			29.25	14.62	75.2
		177	14.75			
April	12			29.83	14.91	80.5
		181	15.08			
May	13			30.50	15.25	85.2
		185	15.42			
June	15			31.17	15.58	96.3
		189	15.75			
July	15			31.67	15.83	94.7
		191	15.92			
Aug.	17			31.92	15.96	106.5
		192	16.00			
Sept.	18			32.17	16.08	111.9
		194	16.17			
Oct.	20			32.59	16.29	122.8
		197	16.42			
Nov.	22			33.09	16.54	133.0
		200	16.67			
Dec.	24			33.59	16.79	142.9
		203	16.92			
2010						
Jan.	12			34.25	17.12	70.1
		208	17.33			
Feb.	13			34.91	17.45	74.5
		211	17.58			
March	13			35.41	17.70	73.4
		214	17.83			
April	15			35.83	17.91	83.7
		216	18.00			
May	16			36.17	18.08	88.5
		218	18.17			
June	18			36.42	18.21	98.8
		219	18.25			
July	20					
Aug.	20					
Sept.	21					
Oct.	22					
Nov.	24					
Dec.	25					

COMPUTATION OF SEASONAL INDICES

	2007	2008	2009	2010	Median	Seasonal index
Jan.		69.5	70.0	70.1	70.0	70.28
Feb.		70.4	83.3	74.5	74.5	74.80
March		78.1	75.2	73.4	75.2	75.50
April		86.0	80.5	83.7	83.7	84.03
May		94.4	85.2	88.5	88.5	88.85
June		96.0	96.3	98.8	96.3	96.38
July	102.8	103.7	94.7		102.8	103.21
Aug.	108.8	103.7	106.5		106.5	106.93
Sept.	109.4	103.7	111.9		109.4	109.84
Oct.	116.3	111.3	122.8		116.3	116.77
Nov.	136.1	126.0	133.0		133.0	133.53
Dec.	137.5	140.0	142.9		140.0	140.56
					1,196.2	1,200.68*

It should be noted that there are only three values for each month since the moving average failed to provide averages for the first half of 2007 and the iast half of 2010. Median has been used to average the figures given for the individual months. The sum of 12 values obtained is 1,196.2. It is necessary, therefore, to make an adjustment so that the total is 1,200. The adjustment is done by multiplying the average (median) values by $\frac{1200}{1196.2} = 1.003$. The final result thus obtained gives us the seasonal indices. The interpretation of this index is very simple. Typical April sales are 84.03 per cent of those of the average month, typical November sales are 133.53 per cent of those of the average month, and so on.

Merits and Limitations of the Ratio-to-Moving Average Method

Merits. This method of measuring seasonal variation is considered to be the most satisfactory and as such is more widely used in practice than other methods. The index obtained by the ratio-to-moving average method ordinarily does not fluctuate so much as the index based on straight-line trends. Mathematical methods of avoiding the effects of the business cycle are not usually needed, for the 12-month moving average follows the cyclical course of the actual data quite closely. Therefore, the index ratios are often more representative of the data from which they are obtained than in the ratio-to-trend method. Also ratio-to-moving average method allows for greater flexibility.

Limitations. However, one drawback of this method is that seasonal indices cannot be obtained for each month for which data are available. When a 12-month moving average is taken, six months in the beginning and six months in the end are left out for which we cannot calculate seasonal indices.

4. Link Relatives Method

Among all the methods of measuring seasonal variation, link relatives method is the most difficult one. When this method is adopted, the following steps are taken to calculate the seasonal variation indices:

1. Calculate the link relatives of the seasonal figures. Link relatives are calculated by dividing the figure of each season** by the figure of immediately preceding season and multiplying it by 100.

*The difference is due to approximation.

**The word season refers to the time period. In case of monthly data, season would refer to a month and in case of quarterly data to a quarter.

$$\frac{\text{Current season's figure}}{\text{Previous season's figure}} \times 100$$

These percentages are called link relatives since they link each month (or quarter or other time period) to the preceding one.

2. Calculating the average of the link relatives for each season. While calculating average we might take arithmetic average but the median is probably better. The arithmetic average would give undue weight to extreme cases which were not due primarily to seasonal influences.

3. Convert these averages into chain relatives on the base of the first season.

4. Calculate the chain relatives of the first season on the base of the last season. There will be some difference between the chain relative of the first season and the chain relative calculated by the previous method, This difference will be due to the effect of long-term changes. It is, therefore, necessary to correct these chain relatives.

5. For correction, the chain relative of the first season calculated by first method is deducted from the chain relative (of the first season) calculated by the second method. The difference is divided by the number of seasons. The resulting figure multiplied by 1, 2, 3 (and so on) is deducted respectively from the chain relatives of the 2nd, 3rd, 4th (and so on) seasons. These are correct chain relatives.

6. Express the corrected chain relatives as percentage of their averages. These provide the required seasonal indices by the method of link relatives.

The following example will illustrate the process :

Illustration 16. Apply method of link relatives to the following data and calculate seasonal indices.

QUARTERLY FIGURES

Quarter	*2006*	*2007*	*2008*	*2009*	*2010*
I	6.0	5.4	6.8	7.2	6.6
II	6.5	7.9	6.5	5.8	7.3
III	7.8	8.4	9.3	7.5	8.0
IV	8.7	7.3	6.4	8.5	7.1

Solution. CALCULATION OF SEASONAL INDICES BY METHOD OF LINK RELATIVES

Year	*Quarter* I	II	III	IV
2006	—	108.3	120.0	111.5
2007	62.1	146.3	106.3	86.9
2008	93.2	95.6	143.1	68.8
2009	112.5	80.6	129.3	113.3
2010	77.6	110.6	109.6	88.8
Arithmetic Average	$\frac{345.4}{4} = 86.35$	$\frac{541.4}{5} = 108.28$	$\frac{608.3}{5} = 121.66$	$\frac{469.3}{5} = 93.86$
Chain relative	100	$\frac{100 \times 108.28}{100} = 108.28$	$\frac{121.66 \times 108.28}{100} = 131.73$	$\frac{93.86 \times 131.73}{100} = 123.64$
Corrected Chain relative	100	108.28–1.675 = 106.605	131.73–3.35 = 128.38	123.64–5.025 = 118.615
Seasonal Indices	100	$\frac{106.605}{113.4} \times 100 = 94.00$	$\frac{128.38}{113.4} \times 100 = 113.21$	$\frac{118.615}{113.4} \times 100 = 104.60$

In the above table the correction factor has been calculated as follows :
Chain relative of the first quarter
(on the basis of first quarter) = 100
Chain relative of the first quarter
(on the basis of the last quarter)

$$\frac{86.35 \times 123.64}{100} = 106.76$$

The difference between these chain relatives = 106.76 – 100 = 6.76

$$\text{Difference per quarter} = \frac{6.7}{4} = 1.675$$

Adjusted chain relatives are obtained by subtracting 1 × 1.675; 2 × 1.675 ; 3 × 1.675 from the chain relatives of the 2nd, 3rd and 4th quarters, respectively.

Seasonal variation indices have been calculated as follows:

$$\frac{100 + 106.605 + 128.38 + 118.615}{4} = \frac{453.6}{4} = 113.4$$

$$\text{Seasonal variation index} = \frac{\text{Correct chain relatives} \times 100}{113.4}$$

Which Method to use. For different methods of measuring seasonal variations have been discussed above. The question now arises which method to adopt in a particular case. The choice will very much depend upon the nature of data and the object of investigation. Amongst all the methods, method of monthly averages is the simplest. But it is a crude method as it assumes that there is no trend component in time series. This method can be used only if seasonal rhythm dominates the data, and trend and cycle are negligible. The method of link relatives was widely used at one time, but its disadvantages seem to outweigh its advantages and it has currently fallen in some disfavour*. On weighing the merits and demerits of ratio-to-trend method and ratio-to-moving average method, one finds that ratio-to-moving average method has several advantages over the ratio-to-trend method. Hence, in general it may be said that because of theoretical and practical advantages, ratio-to-moving average method should be preferred to other methods.

Selecting the Period to Compute Seasonal Indices

In order to simplify the example a period of only 4-5 years was employed to compute the seasonal indexes. In actual fact it is suggested that many more years be included. It is because of the fact that a seasonal index based on a short period is often unduly affected by conditions prevailing during one phase of the business cycle or by powerful random influences. The period should encompass at least one and, if possible, several business cycles. The long span of years offers greater likelihood that irregular and cyclical forces will cancel out or at least have their influence minimized. Ten years is often viewed as a practical minimum. In selecting the period care should be taken to have the period begin and end at the same phase of the business cycle in order to avoid distortions that could result if more years of prosperity than of depression were included.

Average in Computing Seasonals

In each of the methods described for computing seasonal variations, the individual monthly averages were averaged in order to eliminate random influences and any remaining cyclical elements. In one of these example, the average selected was the arithmetic mean and in another the median. This poses the question of the relative merits of these or other averages for the purpose at hand. Since the mean is affected by every item in the series, it should be used when the number of years is large. However, when the period is shorter, the use of the mean is not recommended because extreme items, occasioned by the very random or cyclical factors that the calculation is designed to eliminate,

*Freund and Williams: ***Modern Business Statistics.***

distort its value. The median, on the other hand, is a positional average. As such it is not affected in any way by extreme values, but it may be unduly influenced by the inclusion or exclusion of a year or two in the calculation. A positional mean as suggested by Wessel and Willett avoids the disadvantages of both the mean and the median. It is computed by taking the arithmetic mean of the central items in the series. Suppose the following are the arranged ratios to the moving average for the month of May :

80 85 90 95 96 98 100 102 105 107 112 120

In this case of arithmetic of the mean the middle six items would be employed as the seasonal index. It is obvious that extreme items cannot influence this value and detail of position alone is not significant.

Eliminating Seasonal Influences

The seasonal influences may be removed from time-series data by dividing the actual values for each month by the seasonal index. This adjustment may symbolically be expressed as follows :

$$\frac{T \times S \times C \times I}{S} = T \times C \times I$$

Such adjustments are frequently made for series that manifest significant seasonals when these are being studied for other characteristics.

Illustration 17. From the following data of the production of *XYZ* Co. Ltd. for the year 2010, remove the seasonal influence :

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Production	90	100	110	112	118	150	125	118	110	107	102	99
Seasonal Index	87.0	95.2	102.4	104	106	115	110	103.6	99	105	108	85

Month	*Production*	*Seasonal index*	*Adjusted production*
Jan.	90	87.0	103.45
Feb.	100	95.2	105.04
March	110	102.4	107.42
April	112	104.0	107.69
May	118	106.0	111.32
June	150	115.0	130.43
July	125	110.0	113.64
Aug.	118	103.6	113.90
Sept.	110	99.0	111.11
Oct.	107	105.0	101.90
Nov.	102	108.0	94.44
Dec.	99	85.0	116.47

Uses and Limitations of Seasonal Index

Uses. A seasonal index may be used either analytically or synthetically. Analytically a seasonal index is employed to adjust original data in order to yield deseasonalised data that permit the study of short-run fluctuations of a series not associated with seasonal variations. The procedure of adjusting data for seasonal variations is a simple one. It involves merely the division of each of the original observations by the appropriate seasonal index for that month, *i.e.*,

$$TCI = \frac{TSCI}{S}$$

Synthetically, seasonal index is extremely useful in planning sales of production for specific periods. For example, if a firm expects to sell Rs. 36,00,000 worth of goods during the forthcoming year, average monthly sales of Rs. 3,00,000 is anticipated. If, however, the volume of sales is subjected to seasonal fluctuations, the actual monthly values will deviate significantly from this average. Should the seasonal index for September be 120, the firm can expect sales of Rs. 3,60,000 during that month, in comparison of an index of 90 for July would lead them to anticipate sales only Rs. 2,70,000.

Forecasts for future periods are frequently made by combining what is known about trend and seasonal elements. First, the trend ordinate for a given month is computed. Then this ordinate is multiplied by the seasonal index for that month. For example, if the equation for the trend of a company's sales is $y = 30{,}000 + 250x$, where x represents 1 month and has a value of 0 in December 2009, and the seasonal index for May is +10 the sales for October 2010 may be estimated as follows. In October 2010 the value of x will be +10 and the trend ordinate will, therefore, be 32,500. When this figure is multiplied by 1.1, the estimated sales will be Rs. 35,750. Though this type of forecast ignores cyclical and random influences it is found to be highly useful in practice. By special price and advertising policies a producer confronted with a strong seasonal demand for his product may try to stabilise sales by encouraging off-season consumption.

The most promising solution for seasonality is diversification. It benefits not only the firm but also society at large. Whenever diversification is possible, real costs of seasonal variations can be reduced or even eliminated. Diversification involves the development of production lines having complementary seasonal movements. While some expand seasonally, others contract. Consequently, labour and facilities can be transferred from one line to another as seasonal changes take place. However, diversification is possible only in those line of production that have approximately the same labour and equipment requirements.

Limitations. While making use of seasonal indexes in business and economic problems, the following precautions should be kept in mind:

1. No technique can measure seasonal variations precisely. The various methods of measuring seasonal variations are based on rather unrealistic assumption that the seasonals are changing in some regular and systematic pattern.

2. In developing seasonal index we obtain a series of measures—measure for January, measure for February, and so forth—each of which generally differs from 100. However, we must remember that these measures are only rough estimates. Hence, if we obtain a seasonal index in which the values are all close to 100—for example, if the index values for the consecutive months are 102, 99, 103, 98, etc., it may well be that no real monthly seasonal variation exists in the series and that the small differences from 100 are only due to random influences or imperfect measurement.

3. Even if the computer index of seasonal variation indicates a pronounced pattern, it may have no significance for a particular year. It must be remembered that any seasonal index of the type we have described represents an average pattern during a number of years. If the pattern of seasonal variation in the series is not a stable one, any average pattern may be a poor representation of the actual seasonal variation taking place during a given year.

III. MEASUREMENT OF CYCLICAL VARIATIONS

Business cycles are perhaps the most important type of fluctuation in economic data. Certainly they have received a lot of attention in economic literature. Despite the importance of business cycles, they are most difficult type of economic fluctuation to measure. This is because successive cycles vary widely in timing, amplitude and pattern, and because the cyclical rhythm is inextricably mixed with irregular

factors. Because of these reasons it is impossible to construct meaningful typical cycle indexes of curves similar to those that have been developed for trends and seasonals. The various methods* used for measuring cyclical variations are :

1. Residual method,
2. Reference cycle analysis method,
3. Direct method, and
4. Harmonic analysis method.

Only the first two methods which are in popular use are discussed below.

Residual Method

Among all the methods of arriving at estimates of the cyclical movements of time series, the residual method is most commonly used. This method consists of eliminating seasonal variation and trend, thus obtaining the cyclical irregular movements. Symbolically,

$$\frac{T \times S \times C \times I}{S} = T \times C \times I$$

and

$$\frac{T \times C \times I}{T} = C \times I$$

The data are usually smoothed in order to obtain the cyclical movements, which are sometimes termed the *cyclical relatives,* since they are always percentages. It is because cyclical, irregular or the cyclical movements remain as residuals that this procedure is referred to as the *residual* method.

Limitations of the Residual Method. If the trend ordinates perfectly depicted the pattern of secular change and if the seasonal index exactly reflected seasonal influence, the residual method would leave values reflecting only cyclical and irregular influences. Because such perfection is rarely encountered, the computed values almost always contain some trend and seasonal elements. This condition will be more or less serious depending on how well or poorly the trend line and the seasonal index represent secular and seasonal forces. If a straight line trend is employed to describe an essentially curvilinear secular movement, figures presumably adjusted for trend will be grossly distorted. The distortion would also occur if the seasonal index were not descriptive of the seasonal pattern at the time in question. Thus the residual method is based on the assumption that trend and seasonals can be accurately measured and therefore be removed at least in large part.*

Reference Cycle Analysis or the National Bureau Method

The National Bureau of Economic Research has developed a different method of analysing cyclical variations which it has used in the study of more than 1,000 specific time series. The method is of value in analysing past cycles only. The National Bureau procedure aims to answer two sets of questions:

(1) Is there in a given series a pattern of change that repeats itself (with more or less variation) in successive cycles in business at large? If so, what are its characteristics?

(2) Is there in a given series a wave movement peculiar to that series? If so, what are its characteristics?

The questions under (1) are concerned with the behaviour of individual series during successive waves of expansion and contraction in the *general economic,* those under (2) relate to periodic or semi-periodic fluctuations in *individual series.* A procedure involving 'reference dates' has been designed by the National Bureau of Economic Research as a device which allows one not only to compare each series with a standard set of dates and to observe the behaviour of individual series during expansion and contraction of general business but also to compare the results for the various individual series.

---

*For detail refer to Croxton and Cowden: ***Applied General Statistics.***

*Wessel and Willett : ***Statistics as Applied to Economics and Business.***

The first step is the selection of the reference dates which are the dates of the peaks and troughs of business cycles. This reference dates which cover a duration of over one year and not over ten or twelve years were chosen after examination of large number of economic time series and after study of the "contemporary" reports of observers of the business scene.

The next step consists of processing the data of the individual series in order to obtain a cyclical pattern for each series for the period between each two successive reference troughs. Each period is the same for all series, enabling one to compare the results for the various series. The processing of each series proceeds as follows:

(1) The data are adjusted for seasonal variation.*

(2) The seasonally adjusted data are divided into reference cycle segments, these segments corresponding to the intervals between adjacent reference troughs.

(3) For each segment, the monthly values are expressed as percentages of the average of the values in the segment. These are "reference cycle relatives". As a result of this step, all series, no matter what the original unit, are in percentage form. This step eliminates inter-cycle trend, since the average of the relatives for each cycle is 100, but it does not eliminate intra-cycle trend. The inclusion of intra-cycle trend is regarded as desirable, since it "helps to reveal and to explain what happens during business cycles".

(4) Each reference cycle segment is broken into nine stages, to correspond to the same nine stages in the business cycle, and the reference cycles relatives are averaged for each of nine stages. The nine stages are identified as follows :

(*i*) The 3 months centered on the initial trough.

(*ii*) The first third of the expansion period.

(*iii*) The second third of the expansion period.

(*iv*) The last third of the expansion period.

(*v*) The 3 months centered on the peak.

(*vi*) The first third of the contraction period.

(*vii*) The second third of the contraction period.

(*viii*) The last third of the contraction period.

(*ix*) The 3 months centered on the terminal trough.

The nine-stage average for each reference cycle segment serves to reduce the erratic movement in a series and gives a reference cycle pattern for the particular series under consideration.

Although the National Bureau method of cycle analysis may seem more complicated and cumbersome than the residual technique, it has proved to be the simplest and most accurate way of comparing the cyclical variations of individual series with those of general business. In addition, it is free of errors that might be introduced were secular trend improperly estimated. The latter advantage is indeed significant when series whose trend patterns are not clear are under analysis. Its principal shortcoming is found in the fact that, because no cycle can be studies in this way until it is completed, the method cannot be applied to current data.

MEASUREMENT OF IRREGULAR VARIATIONS

The irregular component in a time series represents the residue of fluctuations after trend cyclical and seasonal movements have been accounted for. Thus, if the original data is divided by T, S and C; we

*Trend influences are not removed under this method.

get I, *i.e.*, $\left(\frac{TSCI}{TSC} = I\right)$. In practice, the cycle itself is so erratic and is so interwoven with irregular movements that it is impossible to separate them. In the analysis of a time series into its component fluctuations, therefore, trend and seasonal movements are usually measured directly, while cyclical and irregular fluctuations are left altogether after the other elements have been removed.

Selecting the Appropriate Forecasting Technique

Numerous forecasting techniques with varying degrees of complexity have bee.ı devised during the last few decades. The problem very often is that of selecting the best one in a prticular situation. The following are some of the important factors that affect the decision about the appropriate forecasting technique :

(*i*) **The Time Horizon.** The period of time over which a decision will have an impact and for which the manager must plan clearly affects the choice of forecasting techniques. Time horizons are generally divided into four heads—immediate term (less than one month) ; short term (one to three months), medium term (three months to two years) and long term (more than two years). Though the exact length of time that may describe each of these categories may vary from company to company, some set of guidelines is necessary so that the forecast will be appropriate for the planning horizon used by the decision maker. Some techniques are appropriate for forecasting only one or two periods in advance whereas others can be used for several periods. Generally speaking, quantitative methods of forecasting are more appropriate for intermediate and shorter term whereas qualitative methods of forecasting are used much more for longer-term forecasts.

The time allowed for preparing the forecast must also be considered. Some methods are much time-consuming and may have to be ignored on the ground only even though they may ensure better results.

(*ii*) **Number of items.** In a situation in which only a single item is being forecast the rules used in preparing that forecast can be much more detailed and complex. But if forecasts are to be made for hundreds or thousands of products, it would be better to develop simple decision rules that can be applied mechanically to each of the items.

(*iii*) **Details required.** In selecting a forecasting technique for a specific situation, one must be aware of the level of detail that will be required for that forecast to be useful in making decisions.

(*iv*) **The pattern of data.** The forecasting methods generally involve an assumption of the type of pattern found in the data. For example, some series depict a seasonal as well as a trend pattern whereas others consist of an average value with random fluctuation surrounding it. Since different forecasting methods vary in their ability to identify different patterns, it is important to match the presumed pattern in the data with the appropriate technique.

(*v*) **Type of model.** In addition to assuming some basic underlying pattern in the data, more forecasting methods also assume some model of the situation being forecast. The model may be a casual model that represents the forecast as being dependent on the occurrence of a number of different events and one may have to use regression or correlation analysis or it may be a series in which time is viewed as the important element in determining changes in the pattern or it may be a mixed model combining in itself a number of different models. The assumptions underlying different models are different and the capabilities of different models in various decision-making situations also vary.

(*vi*) **Cost.** Cost is a very important factor which has to be taken into account while selecting appropriate forecasting technique. The variation in cost affects the attractiveness of different methods for different situations. Generally, four elements of cost that should be considered in the application of forecasting procedure are development, storage, actual operation and opportunity in terms of other techniques.

(*vii*) **Accuracy desired.** The choice of appropriate method of forecasting would also depend upon the accuracy desired. In some situations, variation anywhere between plus and minus 10% may be sufficient for the purpose whereas in other cases a variation as low as 2% to 3% may spell disaster for the company.

(*viii*) **Ease of Application.** The forecasting technique may vary from simple to highly complicated one. Since the manager is held responsible for his decisions, he should not base them on forecasts that he does not understand or in which he has no confidence. Hence, in addition to meeting the requirements of the situation, the forecasting technique must fit with the particular manager who will use the forecast. The manager will have to use his own judgment in order to be able to evaluate and mark selections in his own situation. If a manager can use a more straightforward and less expensive forecasting method rather than the most sophisticated technique available and still achieve the required level of accuracy, he should do so.

It should be borne in mind that a continuous review process must be established in order that the forecast may be compared with the actual results and improvements, and perhaps changes in the technique itself can be made. If no evaluation is made, the manager may feel disappointed with the technique used and may begin to discount it. However, at this stage there may be no record of actual *versus* forecast and other evaluation measures and thus it may be difficult, if not impossible, to determine where improvements are required.

Cautions while using Forecasting Techniques

Forecasting business conditions is a complex task which cannot be accomplished with exactness. The economic, social and political forces which shape the future are many and varied; their relative importance change almost constantly. It is obvious, therefore, that statistical methods cannot claim to be able to make the uncertain future certain—after all, forecasters are not prophets. It does not follow from this disclaimer that statistical methods have nothing to contribute to business forecasting. The choice is not between forecasting and not forecasting, because the lack of a forecast implies a dangerous type of forecast, the mere warning of a possibility of a change is better than no warning at all, as is wisely said "*forewarned is forearmed*".

No matter what method of forecasting is used it is essential that the forecasts be checked by the judgment of individual who is familiar with the business. While it is true that the use of statistical data is an attempt to substitute facts for subjective judgment it does not mean that knowledge gained through experience in a given situation should be ignored in favour of quantitative data. It is particularly important to take into consideration any specific plans of the business that might affect the pattern of sales in relation to indicators used for forecasting. More successful forecasting will result by combining with statistical forecasting the judgment and knowledge of current business trends.

Also it is important to emphasise that any forecast should be reviewed frequently and revised in the light of the most recent information. Forecasting is not a one-shot operation. To be effective it requires continuous attention. Unanticipated developments will often change our picture of the future, or at least clarify it. In terms of any original decisions and actions that have been taken, this rule implies continuous modification where possible. The technique of flexible budgets has been developed to permit the revision of the budget estimates, and everyone dealing with forecasts should be alert to the need for constantly checking to see if anything has happened to change the outlook. Keeping accurately informed about the current level of business is probably the simplest insurance that can be secured against making wrong decisions regarding the future.

Last but not the least it should be kept in mind that as is the case with any method employed to forecast the future, the prediction is no better than the data used no matter how elaborate or complicated the mathematical procedure. As one expert has stated, "It is far better to be approximately correct than precisely wrong." Too often the mathematically-oriented person forgets this point in his zeal to apply his newly discovered tools.

MISCELLANEOUS ILLUSTRATIONS

Illustration 18. Fit a straight line trend to the following time series data :

Year	:	2006	2007	2008	2009	2010
Sale of sugar (in m tonnes)	:	80	90	92	83	94

Eliminate trend from the series. What components are left over ?

Solution. FITTING STRAIGHT LINE TREND

Year	*Sales* Y	X	XY	X^2	Y_c	$(Y-Y_c)$
2006	80	–2	–160	4	83.6	– 3.6
2007	90	–1	–90	1	85.7	+ 4.3
2008	92	0	0	0	87.8	+ 4.2
2009	83	+1	+83	1	89.9	– 6.9
2010	94	+2	+188	4	92.0	+ 2.0
$N = 5$	$\Sigma Y = 439$	$\Sigma X = 0$	$\Sigma XY = 21$	$\Sigma X^2 = 10$		

$$Y_c = a + bX$$

$$a = \frac{\Sigma Y}{N} = \frac{439}{5} = 87.8; \quad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{21}{10} = 2.1$$

Hence $Y = 87.8 + 2.1X$

$Y_{2006} = 87.8 + 2.1\,(-2) = 87.8 - 4.2 = 83.6$; and similarly other trend values are computed.

After eliminating trend what is left is the effect of seasonal, cyclical and irregular variations.

Illustration 19. Below are given the figures of production (in million tonnes) of a cement factory :

Year	:	2001	2003	2004	2005	2006	2007	2010
Production (in m. tonnes)	:	77	88	94	85	91	98	90

(*i*) Fit a straight line trend by the 'least squares method' and tabulate the trend values.

(*ii*) Eliminate the trend. What components of the Time Series are thus left over ?

(*iii*) What is monthly increase in the production of cement ?

Solution. FITTING STRAIGHT LINE TREND BY THE METHOD OF LEAST SQUARES

Year	*Production (in m. tonnes)* Y	*Deviations from 2005* X	XY	X^2	*Trend values* Y_c
2001	77	–4	–308	16	83.299
2003	88	–2	–176	4	86.051
2004	94	–1	–94	1	87.427
2005	85	0	0	0	88.803
2006	91	+1	+91	1	90.179
2007	98	+2	+196	4	91.555
2010	90	+5	+450	25	95.683
$N = 7$	$\Sigma Y = 623$	$\Sigma X = 1$	$\Sigma XY = 159$	$\Sigma X^2 = 51$	

(*i*) The equation of the straight line trend is $Y_c = a + bX$. Since ΣX is not zero, we will solve the two normal equations

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

$$623 = 7a + b$$

$$159 = a + 51b$$

Multiplying eq. (*ii*) by 7 and subtracting from eq. (*i*), we get $b = 1.376$

Putting the value of b in eq. (*i*)

$7a + 1.376 = 623$ or $7a = 621.624$ or $a = 88.803$

Hence $Y = 88.803 + 1.376X$ is the equation of the straight line

$Y_{2001} = 88.803 + 1.376\ (-4) = 88.803 - 5.504 = 83.299$

$Y_{2003} = 88.803 + 1.376\ (-2) = 88.803 - 2.752 = 86.051$

$Y_{2004} = 88.803 + 1.376\ (-1) = 88.803 - 1.376 = 87.427$

$Y_{2005} = 88.803$

$Y_{2006} = 88.803 + 1.376\ (+1) = 88.803 + 1.376 = 90.179$

$Y_{2007} = 88.803 + 1.376\ (+2) = 88.803 + 2.752 = 91.555$

$Y_{2010} = 88.803 + 1.376\ (+5) = 88.803 + 6.88 = 95.683$

(*ii*) After eliminating the trend we are left with seasonal, cyclical and irregular variations.

(*iii*) Monthly increase in the production of cement shall be given by

$$\frac{b}{12} = \frac{1.376}{12} = 0.115$$

Illustration 20. The sale of a commodity (in tonnes) varied from January 2010 to December, 2010 in the following manner :

280	300	280	280	270	240
230	230	220	200	210	200

Fit a trend line by the method of semi-averages.

Solution. FITTING TREND LINE BY METHOD OF SEMI-AVERAGES

Month	*Sales (in tonnes)*		*Month*	*Sales (in tonnes)*	
January	280		July	230	
February	300	1,650 (Total) of first six months	August	230	1,290 (Total) of last six months
March	280		September	220	
April	280		October	200	
May	270		November	210	
June	240		December	200	

$$\text{Average of the first half} = \frac{1650}{6} = 275 \text{ tonnes.}$$

$$\text{Average of the second half} = \frac{1290}{6} = 215 \text{ tonnes.}$$

These two figures, namely, 275 and 215, shall be plotted at the middle of their respective periods, *i.e.*, middle of March-April 2010 and that of September-October 2010. By joining these two points we get a trend line which describes the given data.

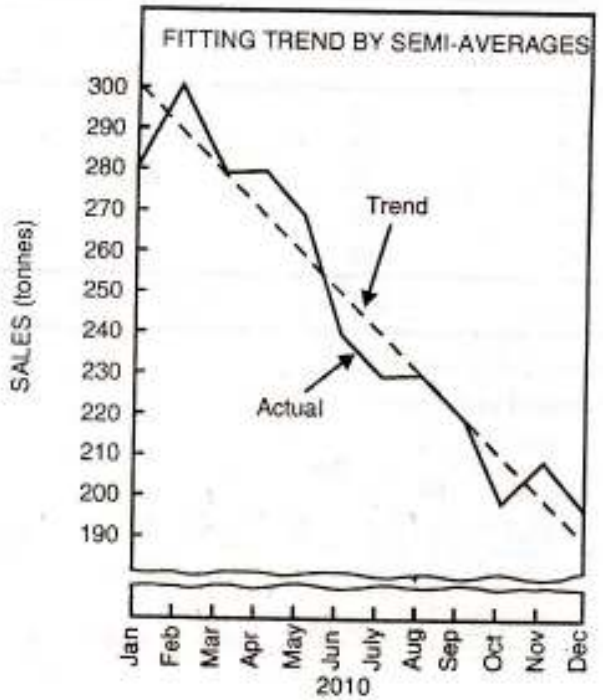


Illustration 21. (*i*) Given the trend equation :

$$Y_c = 35 + 5X + 3X^2$$

(Origin : 2004, X unit = 1 year) change the origin of the equation to 2010.

(*ii*) Given the equation $Y_c = 10(1.5)^X$

(Origin : 2004, X unit = 1 year)

Shift the origin forward by two years.

(*iii*) The trend of the annual sales of an Aluminium Company is described by the following equation :

$$Y_c = 12 + 0.7X$$

(Origin : 2004, X unit = 1year and Y unit = annual production)

Shift the origin to January, 2010 and write the equation on monthly basis.

Solution. (*i*) To shift the origin from 2004 to 2010, *i.e.*, 6 years forward, put $X = 6$ in the equation.

$$Y_c = 35 + 5(6) + 3(6)^2 = 35 + 30 + 108 = 173$$

The new trend equation is

$$Y_c = 173 + 5X + 3X^2$$

(Origin : 2004, X unit = 1 year)

(*ii*) For shifting the origin forward by two years, we put $X + 2$ instead of X in the given equation.

The new trend equation is

$$Y_c = 10(1.5)^{X+2}$$
$$= 10\,(1.5)^2\,(1.5)^X = 22.5\,(1.5)^X$$

(*iii*) To shift the origin to January, 2010, we should subtract $1/2b$ from trend value of July, 2010.

$1/2b = 0.7/2 = 0.35$; therefore, $a = 12 - 0.35 = 11.65$

The new trend equation is

$$Y_c = 11.65 + 0.7X$$

(Origin : January, 2010, X unit = 1 year)

To obtain the equation on monthly basis, we divide the value of a by 12 and value of b by 144. The new trend equation on monthly basis can be written as :

$$Y_c = \frac{11.65}{12} + \frac{0.7}{144}\ X = 0.971 + .0049\ X$$

(Origin : January 2010, X unit = 1 month).

Illustration 22. Fit a parabolic curve of the second degree to the data given below and estimate the value for 2012 and comment on it.

Year	:	2006	2007	2008	2009	2010
Sales (in '000 Rs.)	:	10	12	13	10	8

Solution. COMPUTATION OF SECOND DEGREE PARABOLA

Year	*Sales ('000 Rs.)* Y	X	XY	X^2	X^2Y	X^4
2006	10	–2	–20	4	40	16
2007	12	–1	–12	1	12	1
2008	13	0	0	0	0	0
2009	10	+1	+10	1	10	1
2010	8	+2	+16	4	32	16
$N = 5$	$\Sigma Y = 53$	$\Sigma X = 0$	$\Sigma XY = -6$	$\Sigma X^2 = 10$	$\Sigma X^2Y = 94$	$\Sigma X^4 = 34$

The equation of the second parabola is : $Y = a + bX + cX^2$

The values of a, b and c can be determined as follows :

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{-6}{10} = -0.6$$

$$c = \frac{N\Sigma X^2Y - \Sigma X^2 \Sigma Y}{N\Sigma X^4 - (X^2)^2} = \frac{5 \times 94 - 10 \times 53}{5 \times 34 - (10)^2} = -0.857$$

$$a = \frac{\Sigma Y - c\Sigma X^2}{N} = \frac{53 - (.857 \times 10)}{5} = 8.886$$

Thus, $Y = 8.886 - 0.6X - 0.857X^2$

For 2012, X would be 4 $Y_{2012} = 8.886 - 0.6(4) - 0.857(4)^2$

$= 8.886 - 2.4 - 13.712 = -7.226.$

The expected sales for 2012 comes to be negative. The second degree parabola does not seem to describe the data well.

Illustration 23. Assuming that trend is absent, determine if there is any seasonality in the data given below :

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2007	3.7	4.1	3.3	3.5
2008	3.7	3.9	3.6	3.6
2009	4.0	4.1	3.3	3.1
2010	3.3	4.4	4.0	4.0

What are the seasonal indices for various quarters ?

Solution. COMPUTATION OF SEASONAL INDICES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2007	3.7	4.1	3.3	3.5
2008	3.7	3.9	3.6	3.6
2009	4.0	4.1	3.3	3.1
2010	3.3	4.4	4.0	4.0
Total	14.7	16.5	14.2	14.2
Average	3.675	4.125	3.55	3.55
Seasonal index	98.7	110.8	95.3	95.3

Notes for calculating seasonal index

The average of averages $= \frac{3.675+4.125+3.55+3.55}{4} = \frac{14.900}{4} = 3.725$

$$\text{Seasonal Index} = \frac{\text{Quarterly average}}{\text{General average}} \times 100$$

Seasonal Index for first quarter : $\frac{3.675}{3.725} \times 100 = 98.7$

Seasonal Index for second quarter : $\frac{4.125}{3.725} \times 100 = 110.7$

Seasonal Index for the third quarter : $\frac{3.55}{3.725} \times 100 = 95.3$

Seasonal Index for the fourth quarter : $\frac{3.55}{3.725} \times 100 = 95.3$

Illustration 24. Given below are the figures of production of a sugar factory :

Year	:	2004	2005	2006	2007	2008	2009	2010
Production (m. tonnes)	:	40	45	46	42	47	49	46

Fit a straight line trend by the method of least squares and estimate the value for 2014

Solution. FITTING STRAIGHT LINE BY METHOD OF LEAST SQUARES

Year	*Production (m. tonnes)* Y	*Deviations from 2007* X	XY	X^2
2004	40	–3	– 120	9
2005	45	–2	– 90	4
2006	46	–1	– 46	1
2007	42	0	0	0
2008	47	+1	+ 47	1
2009	49	+2	+ 98	4
2010	46	+3	+ 138	9
$N = 7$	$\Sigma Y = 315$	$\Sigma X = 0$	$\Sigma XY = 27$	$\Sigma X^2 = 28$

$$Y_c = a + bX$$

$$a = \frac{\Sigma Y}{N} = \frac{315}{7} = 45;\ b = \frac{\Sigma XY}{\Sigma X^2} = \frac{27}{28} = 0.964$$

$$Y = 45 + 0.964X$$

$$Y_{2014} = 45 + .964\,(7) = 45 + 6.748 = 51.748$$

Thus, the estimated production for 2014 is 51.748 m. tonnes.

Illustration 25. Fit a straight line trend by the method of least squares to the data given below :

Year	:	2004	2005	2006	2007	2008	2009	2010
Sales (m. tonnes)	:	9	11	13	12	14	15	17

Estimate the likely sales for the year 2013.

Solution. CALCULATION OF STRAIGHT LINE TREND BY METHOD OF LEAST SQUARES

Year	*Sales (m. tonnes)* Y	X	XY	X^2
2004	9	–3	–27	9
2005	11	–2	–22	4
2006	13	–1	–13	1
2007	12	0	0	0
2008	14	+1	+14	1
2009	15	+2	+30	4
2010	17	+3	+51	9
$N = 7$	$\Sigma Y = 91$	$\Sigma X = 0$	$\Sigma XY = 33$	$\Sigma X^2 = 28$

The equation of the straight line trend is : $Y_c = a + bX$

Since $\Sigma X = 0$, $a = \frac{\Sigma Y}{N} = \frac{91}{7} = 13$; $b = \frac{\Sigma XY}{\Sigma X^2} = \frac{33}{28} = 1.179$

Hence $Y = 13 + 1.179X$; For 2011, X would be + 6.

$$Y_{2013} = 13 + 1.179(6) = 13 + 7.074 = 20.074$$

Therefore, the estimated sales for the year 2013 is 20.074 m. tonnes.

Illustration 26. The seasonal indices of a commodity manufactured by a company for four quarters of a year are respectively 100, 90, 80, and 130. If the total sale in the first quarter is worth Rs. 25,000, how much worth of sale is expected during the whole year ?

Solution. EXPECTED SALES IN VARIOUS QUARTERS

Quarter	*Seasonal index*	*Estimated sales (Rs.)*
I	100	25,000
II	90	22,500
III	80	20,000
IV	130	32,500

Estimated sales for 2nd Qtr. $= \frac{\text{Figure for 1st Qtr.}}{\text{S.I. for Ist Qtr.}} \times \text{S.I. for 2nd Qtr.}$

$$= \frac{25{,}000 \times 90}{100} = 22{,}500$$

Estimated sales for 3rd Qtr. $= \frac{\text{Figure for 1st Qtr.}}{\text{S.I. for Ist Qtr.}} \times \text{S.I. for 3rd Qtr.}$

$$= \frac{25{,}000 \times 80}{100} = 20{,}000$$

Estimated sales for 4th Qtr. $= \frac{25{,}000 \times 130}{100} = 32{,}500.$

Illustration 27. Fit a straight line trend by the method of least squares to the following data on sales (Rs. in lakh) for the period 2003-2010.

Year	:	2003	2004	2005	2006	2007	2008	2009	2010
Sales (Rs. lakh)	:	76	80	130	144	138	120	174	190

Also :

(*a*) Calculate the trend values from 2003 to 2010.

(*b*) What will be predicted sales for 2013, assuming that the same rate of change continues.

Solution. FITTING STRAIGHT LINE TREND BY THE METHOD OF LEAST SQUARES

Year	*Sales (Rs. Lakh) Y*	*Deviations from 2006.5 × 2*	*X*	*XY*	X^2	Y_c
2003	76	–3.5	–7	– 532	49	80.169
2004	80	–2.5	–5	– 400	25	94.835
2005	130	–1.5	–3	–390	9	109.501
2006	144	–0.5	–1	–144	1	124.167
2007	138	+0.5	+1	+138	1	138.833
2008	120	+1.5	+3	+360	9	153.499
2009	174	+2.5	+5	+870	25	168.165
2010	190	+3.5	+7	+1330	49	182.831
$N = 8$	$\Sigma Y = 1052$		$\Sigma X = 0$	$\Sigma XY = 1232$	$\Sigma X^2 = 168$	$\Sigma Y_c = 1052$

Since $\Sigma X = 0$, $a = \frac{\Sigma Y}{N} = \frac{1052}{8} = 131.5$; and $b = \frac{\Sigma XY}{\Sigma X^2} = \frac{1232}{168} = 7.333$*

Hence $Y = 131.5 + 7.333X$

For the year 2013, X would be 13, therefore $Y_{2013} = 131.5 + 7.333\ (13) = 131.5 + 95.329 = 226.829$ = Rs. 2,26,829.

Illustration 28. Fit a straight line trend for the following data and find the trend values. Estimate the sales for 2016.

Year	:	2004	2005	2006	2007	2008	2009	2010
Sales (Rs. lakh)	:	33	35	60	67	68	82	90

Solution. FITTING STRAIGHT LINE TREND

Year	*Sales Y*	*X*	*XY*	X^2	Y_c
2004	33	–3	–99	9	32.893
2005	35	–2	–70	4	42.643
2006	60	–1	–60	1	52.393
2007	67	0	0	0	62.143
2008	68	+1	+68	1	71.893
2009	82	+2	+164	4	81.643
2010	90	+3	+270	9	91.393
$N = 7$	$\Sigma Y = 435$	$\Sigma X = 0$	$\Sigma XY = 273$	$\Sigma X^2 = 28$	$\Sigma Y_c = 435$

The equation of the straight line trend is :

$$Y = a + bX$$

$$a = \frac{\Sigma Y}{N} = \frac{435}{7} = 62.143\ ;\quad b = \frac{\Sigma XY}{X^2} = \frac{273}{28} = 9.75$$

Hence,

$$Y = 62.143 + 9.75X$$

$$Y_{2004} = 62.143 + 9.75(-3) = 62.143 - 29.25 = 32.893$$

$$Y_{2005} = 62.143 + 9.75(-2) = 42.643, \text{ etc.}$$

For 2016, X shall be + 9; $Y_{2016} = 62.143 + 9.75(9) = 62.143 + 87.75 = 149.893$

The estimated sales for the year 2016 is Rs. 149.893 lakh.

*The calculated value of *b*, *i.e.*, 7.333 is multiplied by 2, *i.e.* (7.333 × 2 = 14.666) to obtain yearly change.

Illustration 29. Fit a straight line trend by the method of least squares to the following data :

Year	:	2004	2005	2006	2007	2008	2009	2010
Production of steel (m. tonnes)	:	12	10	14	11	13	15	16

Calculate the trend values and estimate the likely production for the year 2017. Interpret the values of a and b.

Solution.

CALCULATION OF TREND VALUES

Year	*Production (m. tonnes)* Y	*Deviations from 2007* X	XY	X^2	*Trend values* Y_c
2004	12	–3	–36	9	10.75
2005	10	–2	–20	4	11.50
2006	14	–1	–14	1	12.25
2007	11	0	0	0	13.00
2008	13	+1	+13	1	13.75
2009	15	+2	+30	4	14.50
2010	16	+3	+48	9	15.25
$N = 7$	$\Sigma Y = 91$	$\Sigma X = 0$	$\Sigma XY = 21$	$\Sigma X^2 = 28$	$\Sigma Y_c = 91$

The equation of the straight line trend is : $Y = a + bX$.

Since $$\Sigma X = 0, a = \frac{\Sigma Y}{N} = \frac{91}{7} = 13 ; b = \frac{\Sigma XY}{X^2} = \frac{21}{28} = 0.75$$

Hence, $$Y = 13 + 0.75X ; \text{ For 2004 : } X = -3$$

Estimated production for the year 2004 :

$$Y = 13 + 0.75(-3) = 13 - 2.25 = 10.75$$

Estimated value for 2005 = $13 + 0.75(-2) = 11.5$, etc.

For the year 2017, X would be +10.

$$Y_{2017} = 13 + 0.75(10) = 13 + 7.5 = 20.5$$

Hence, the estimated production of steel for the year 2017 is 20.5 m. tonnes.

Illustration 30. Fit a second degree parabola, $Y = a + bX + cX^2$, to the following population data of a city :

Year	:	2002	2003	2004	2005	2006	2007	2008	2009	2010
Population (in lakh)	:	5	6	6	7	7	8	9	10	10

(Take the year 2006 as the working origin)

Solution.

FITTING OF SECOND DEGREE PARABOLA

Year	*Population* Y	*Deviations from 2006* X	XY	X^2	X^2Y	X^3	X^4
2002	5	–4	–20	16	80	–64	256
2003	6	–3	–18	9	54	–27	81
2004	6	–2	–12	4	24	–8	16
2005	7	–1	–7	1	7	–1	1
2006	7	0	0	0	0	0	0
2007	8	+1	+8	1	8	+1	1
2008	9	+2	+18	4	36	+8	16
2009	10	+3	+30	9	90	+27	81
2010	10	+4	+40	16	160	+64	256
$N = 9$	$\Sigma Y = 68$	$\Sigma X = 0$	$\Sigma XY = 39$	$\Sigma X^2 = 60$	$\Sigma X^2Y = 459$	$\Sigma X^3 = 0$	$\Sigma X^4 = 708$

$$Y = a + bX + cX^2$$

Since $\Sigma X = 0$, the three normal equations would be

$$\Sigma Y = Na + c\Sigma X^2$$

$$\Sigma XY = b\Sigma X^2$$

$$\Sigma X^2Y = a\Sigma X^2 + c\Sigma X^4$$

Substituting the values from the table

$$68 = 9a + 60c \quad ...(i)$$
$$39 = 60b \quad ...(ii)$$
$$459 = 60a + 708c \quad ...(iii)$$

From eq. (*ii*) $60b = 39$ or $b = 0.65$

Multiplying eq. (*i*) by 20 and eq. (*iii*) by 3

$$1360 = 180a + 1200c$$
$$1377 = 180a + 2124c$$
$$- \quad - \quad -$$
$$17 = 924c \text{ or } c = 0.0184$$

Substituting the value of c in eq. (*i*)

$$68 = 9a + 60\,(0.0184) \text{ or } 9a + 1.104 = 68$$
$$9a = 66.896 \text{ or } a = 7.433$$

Hence $Y = 7.433 + 0.65X + 0.0184X^2$ is the required equation.

Illustration 31. The time series given below shows the number of T.V. sold by a company since 2001.

Years	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
T.V. sold ('000)	42	50	61	75	92	111	120	127	140	138

Find the linear equation that describes the trend in the number of T.V. sold. Also estimate the sale of T.V. in 2012.

Solution. FITTING LINEAR EQUATION

Year	*T.V. Sold ('000)*	*Taking deviations from 2005.5*	*Multiplying deviations by 2* X	XY	X^2
2001	42	–4.5	–9	–378	81
2002	50	–3.5	–7	–350	49
2003	61	–2.5	–5	–305	25
2004	75	–1.5	–3	–225	9
2005	92	–0.5	–1	–92	1
2006	111	+0.5	+1	+111	1
2007	120	+1.5	+3	+360	9
2008	127	+2.5	+5	+635	25
2009	140	+3.5	+7	+980	49
2010	138	+4.5	+9	+1242	81
$N = 10$	$\Sigma Y = 956$		$\Sigma X = 0$	$\Sigma XY = 1978$	$\Sigma X^2 = 330$

The linear equation would be of the form :

$$Y = a + bX$$

The two normal equations shall be

$$\Sigma Y = Na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Since $\Sigma X = 0$

$$a = \frac{\Sigma Y}{N} = \frac{956}{10} = 95.6; \text{ and } b = \frac{\Sigma XY}{\Sigma X^2} = \frac{1978}{330} = 5.994$$

Thus the linear equation is

$$Y_c = 95.6 + 5.994$$

For 2012, X would be +13.

$$Y_{2012} = 95.6 + 5.994\,(13) = 95.6 + 77.92 = 173.52 \text{ thousand}$$

Hence the expected number of T.V. that would be sold in 2012 shall be 1,73,520.

PROBLEMS

1-A : Answer the following questions, each question carries **one** marks:

(*i*) What do you mean by Time series ? Explain the objectives of the analysis of a time series. *(MBA, UPTech. Univ., 2007)*

(*ii*) Write a short note on "Secular trend".

(*iii*) What is seasonal variation ? *(MBA, Madurai-Kamaraj Univ., 2006)*

(*iv*) Write down the most important factors causing seasonal variations. *(MBA, Madurai-Kamaraj Univ., 2003)*

(*v*) What are the normal equations for the straight line $Y = a + bx$?

(*vi*) What are cyclical fluctuations ?

(*vii*) What is Business forecasting ?

(*viii*) Name a few methods of Business forecasting.

(*ix*) How irregular variations are caused ?

(*x*) "Forewarned is Forearmed." Comment.

1-B : Answer the following questions, each question carries **four** marks:

(*i*) What are seasonal variations ? Bring out the factors that cause seasonal variations. *(M.A. Econ., Madras Univ., 2003)*

(*ii*) Narrate the merits and limitations in the use of moving average method. *(M.A. Econ., Madras Univ., 2003)*

(*iii*) Distinguish between the additive and multiplicative models of time series analysis.

(*iv*) Suggest the important adjustments to the mode before analysing time series.

(*v*) Explain the freehand (graphic method) of measuring trend.

2. (*a*) What is Business Forecasting ? Explain clearly its role and limitations.

(*b*) How does analysis of time series helps in making business forecast ?

(*c*) What is forecasting ? Discuss in brief the various theories and methods of business forecasting. *(MBA, Delhi Univ., 1998)*

3. (*a*) Explain clearly the different components into which a time series may be analysed. Explain any method in isolating trend values in a time series. *(MBA, Delhi Univ., MBA, Vikram Univ., 2005)*

(*b*) Explain clearly the meaning of Time Series Analysis. Mention its important components. Explain these components with examples, indicating the importance of each component in business. *(B.Com., Andhra Univ., 2003)*

(*c*) Describe the seasonal variation and cyclical fluctuations in a time series. *(MBA, Anna Univ., 2003)*

4. Explain what do you understand by Time Series. Why is Time Series considered to be an effective tool of forecasting ? *(MBA, BHU, 2002)*

5. (*a*) What is business forecasting ? What are the assumptions on which business forecasts are made? Describe the techniques of forecasting that are commonly employed by big business houses.

(*b*) Explain briefly the additive and multiplicative models of time series. Which of these models is more popular in practice and why?

6. (*a*) Critically examine the various methods that are used for measuring trend. Which method do you think is the best and why ?

(*b*) Explain briefly the different methods of measuring trend. *(MBA, Madras Univ., 2002)*

7. (*a*) How seasonal variations are accounted for in the analysis of Time Series ?

(*b*) What are the common methods in use for eliminating seasonality from a time series data ? Explain any one method taking imaginary figures.

8. Critically examine the various methods that are used for business forecasting. Why is time series considered to be an effective tool for forecasting analysis? Explain.

9. Explain the following terms in the study of time series :

(*i*) Secular trend, (*ii*) Seasonal variation, (*iii*) Cyclical fluctuations.

10. (*a*) What do you understand by 'seasonal variation' in time series data ? Explain their uses.

(*b*) Why do we measure seasonal variations in a time series ?

(*c*) How would you eliminate seasonal influences ? Illustrate with the help of an example.

(*d*) Explain clearly with the help of an illustration how seasonal index is useful in planning sales or production for specific periods. Are there any limitations of seasonal index ?

11. (*a*) Explain the method of Moving Averages in estimating the trend of a time series. What are the disadvantages in using this method ?

(*b*) Explain the concept of 'auto correlation' and its use for time series analysis.Give an example of a single variable with two different time lags. *(MBA, IGNOU 2001)*

12. (*a*) Why do we deseasonalize data ? Explain the ratio-to-moving average method to compile the seasonal index.

(*b*) Explain the following statements :

(*i*) "... the business analyst who uses moving averages to smoothen his data while in the process of trying to discover business cycles, is likely to come up with some non-existent cycles."

(*ii*) "There is nothing sacred in computing seasonal indices by the method of moving average using exclusively monthly data."

(*iii*) "Despite great limitations of statistical forecasting, the forecasting techniques are invaluable to the economist, the businessman and the Government."

13. Suppose you are provided with a given time series data and asked to analyse its general pattern and fluctuations. Describe in detail the steps you would follow in determining the pattern of trend and whether a seasonal and/or a cyclical component contributed to movements in the series.

14. (*a*) (*i*) "A key assumption in the classical method of time series analysis is that each of the component movements in the time series can be isolated individually from a series". Do you agree with this statement ? Does this assumption create any serious limitation to such analysis?

(*ii*) "A 12-month moving average of time series data removes trend and cycle." Do you agree ? Why or why not ?

(*b*) Examine critically the time-lag and the action and the reaction theory of business forecasting. Which of these, in your opinion, is better and why ?

15. Answer the following by a brief statement on each :

(*i*) Why must short-term forecasts be more precise than long-term ones ?

(*ii*) What is the major objective of seasonal analysis ?

(*iii*) What purpose does a seasonal index solve ?

16. (*a*) What is the difference between seasonal fluctuations and cyclical variations in a time series data.

(*b*) Illustrate the historical analogy theory of business forecasting.

(*c*) What is a time series ? What are its components ? Which components of the series is mainly applicable in the following cases ? *(M.B.A., UPTech. Univ., 2005)*

(*i*) A fire in a factory delaying production for one month.

(*ii*) Formation of rocks.

(*iii*) Decrease in the employment in sugar factory during the off-season.

(*iv*) Sale of New Year greeting cards.

(*v*) Fall in death rate due to a advances in science.

(*vi*) An after Deepawali sales in a departmental store.

(*vii*) A need for increased rice production due to a constant increase in population.

17. Indicate three categories of forecasting models and list out five techniques from each category. Describe Delphi technique in detail.

18. (*a*) Critically examine the time-lag and the action and reaction theory of business forecasting. Which of these two is better and why ?

(*b*) While fitting a straight line trend of the type $Y = a + bX$, what is signified by Y, X, a and b ?

19. (*a*) Discuss the role of forecasting as a business tool.

(*b*) How do we manage long-range forecasting and technical change for any organisation ?

(*c*) Write short notes on Delphi method and Historical analogy method for business forecasting.

(*d*) Explain how can we use market surveys as a method of forecasting. Illustrate.

(*e*) Write a lucid note on Box and Jenkin's method of forecasting. *(MBA, Kurukshetra Univ., 2005)*

(*f*) Explain with appropriate example different methods of estimating seasonal variations.

(*g*) What do you understand by Naive (Time series) Quantitative Models of forecasting? *(MBA, Jamia Millia, 2003)*

20. Business today generate a large amount of data continuously. This data may be used to gain information about the system. For one such system, it is known that the relation between variables is non-linear, *i.e.*, in the form $y = ax^b$, where a and b are constants. Use a transformation to make it linear and discuss how would you use the method of least squares to fit a straight line to the transformed linear model.

21. Apply the method of semi-averages for determining trend to the following data and estimate the value for 2015 :

Year	*Sales (Thousand units)*	*Year*	*Sales (Thousand units)*
2005	20	2008	30
2006	24	2009	28
2007	22	2010	32

If the actual figure of sales for 2011 is 35,000 units, how do you account for difference between the figure you obtain and the actual figure given to you ?

22. Plot the following data on graph paper and ascertain trend by the method of semi- averages :

Year	*Sales (million tonnes)*	*Year*	*Sales (million tonnes)*
2004	100	2008	108
2005	120	2009	102
2006	95	2010	112
2007	105		

23. Apply the method of semi-average to depict the long-term tendency of following data and estimate the value for 2013:

Year	*Production (million tonnes)*	*Year*	*Production (million tonnes)*
2003	40	2007	51
2004	44	2008	50
2005	42	2009	54
2006	48	2010	56

24. The following series relate to the profits of a commercial concern for 8 years :

Year	*Profits Rs.*	*Year*	*Profits Rs.*
2003	15,420	2007	26,120
2004	14,420	2008	31,950
2005	15,520	2009	35,360
2006	21,020	2010	35,670

Find the trend of profits. (Assume a three-year cycle and ignore decimals.)

25. Find out the trend values for the following time series of steel production by the method of moving average using 5-point time period for your purpose. State briefly the procedure that would have been adopted if you were to choose a 4-point time period. How does one choose the proper 'period of the moving average' ?

Year	*Production (m. tonnes)*	*Year*	*Production (m. tonnes)*	*Year*	*Production (m. tonnes)*
1993	351	1999	410	2005	502
1994	366	2000	420	2006	540
1995	361	2001	450	2007	557
1996	362	2002	500	2008	571
1997	400	2003	518	2009	586
1998	419	2004	455	2010	612

26. Below are given the figures of production of a sugar factory :

Year	*Production (thousand tonnes)*	*Year*	*Production (thousand tonnes)*
2005	92	2008	92
2006	83	2009	92
2007	94	2010	110

Apply the method of least squares to determine the trend values. Also find out the short-term fluctuations.
[$Y = 95 + 1473X$]

27. Fit a straight line trend by the method of least squares :

Year	*Milk consumption (million litres)*	*Year*	*Milk consumption (million litres)*
2002	102.3	2007	118.7
2003	101.9	2008	124.5
2004	105.8	2009	129.9
2005	112.0	2010	134.8
2006	114.8		

[$Y = 116.1 + 4.3X$]

28. The following are annual profits (in thousands of rupees) of a business firm :

Year	:	2004	2005	2006	2007	2008	2009	2010
Profits (in '000 Rs.)	:	60	72	75	65	80	85	95

(*a*) Use the method of least squares to fit a straight line to the above data.
(*b*) Plot the above figures and draw the line.
(*c*) Also make an estimate of the profits for the year 2011.
[$Y = 76 + 4.86X$; $Y_{2011} = 100.3$]

29. Fit a straight line trend by the method of least squares to the following data :

Year (X)	:	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10
Sales (in lakh Rs.) (Y)	:	278	309	335	378	424	481	521

(Clearly specify the origin and the units of the variables in the trend equation obtained.)
[$Y = 389.43 + 41.5X$]

30. Fit an equation of the type $Y = a + bX + cX^2$ to the following data :

Year	*Production (in '000 tonnes)*	*Year*	*Production (in '000 tonnes)*
2006	70	2009	80
2007	72	2010	90
2008	88		

31. The following table shows the number of letters posted in a particular area during a typical period of four weeks. Assuming that the trend value during the period remains the same, calculate 'seasonal indices' (here daily indices) as percentage of the grand average :

Week	*Sun.*	*Mon.*	*Tue.*	*Wed.*	*Thurs.*	*Fri.*	*Sat.*	*Total*
1	18	161	170	164	153	181	76	923
2	18	165	169	147	148	190	80	917
3	21	162	169	153	155	190	82	932
4	20	165	170	155	150	180	85	925
Total	77	653	678	619	606	741	323	3697

32. The working capital requirements of the *XYZ* Ltd. have been subject to seasonal fluctuations. At the same time, a steady secular advance can be noted. In order to evaluate comprehensively future working capital needs, the treasurer calculated a straight line trend and the seasonal indices. The trend equation is $Y_c = 10{,}000 + 500X$, where X represents a period of 1 month and has a value of 0 in 2010. The seasonal indices are as follows :

Jan.	80	July	125
Feb.	95	Aug.	99
Mar.	90	Sep.	90
Apr.	100	Oct.	102
May.	116	Nov.	105
June	120	Dec.	87

(*a*) Prepare a schedule of estimated working capital requirements for 2010.
(*b*) What factors could cause these estimates to be incorrect?
(*c*) What might be done to compensate for inaccuracies as they become apparent?
(*d*) Would you as a banker have any interest in estimates of this type?

33. In order to find quarterly seasonal indices, first of all the quarter wholesale price for five years (2006-2010) were reduced as percentage of their centred moving averages of four quarters. These percentages are set out in the following table. You are required to calculate the quarterly seasonal indices.

Year	*I*	*II*	*III*	*IV*
2006	—	—	127	134
2007	130	122	122	132
2008	120	120	118	128
2009	126	116	121	130
2010	127	118	—	—

I	*II*	*III*	*IV*
101.0	95.6	98.04	105

34. Consult a copy of Business Statistics in your library and select a series of your own. The series should be for a period of minimum eight years. Then do the following :

(*a*) Compute an appropriate trend line for the series, with first month of series as the origin. Plot the original series and its trend in one diagram.

(*b*) Compute a typical seasonal index for the series by the ratio-to-moving average method. Plot the actual data and moving average figures.

(*c*) What comments can you make on T, S, C and I ?

35. *a*) What do you understand by seasonal fluctuation in time series? Give an example.

(*b*) What are the major uses of seasonal indices in time series analysis? Name four methods by which one can compute a seasonal index from time series data.

(*c*) The sales of a company rose from Rs. 60,000 in the month of August to Rs. 69,000 in the month of September. The seasonal indices for these two months are 105 and 140 respectively. The owner of the company was not at all satisfied with the rise of sale in the month of September by Rs. 9,000. He expected much more because of the seasonal index for the month. What were his estimates of sales for the month of September?

$\left[\text{The expectation was } \frac{60{,}000 \times 140}{105} = \text{Rs. } 80{,}000\right]$

36. (*a*) Given the following trend information :

$Y_c = 60 + 2.48X$ Origin : July 1, 2008

Y in million of rupees X in terms of years

Convert this equation in monthly terms. Be sure your response is in the most practical form.

(*b*) The annual trend equation for the *XYZ* Co. Ltd. is represented by the following :

$Y_c = 468 + 0.20X$ X = years

Y = thousands of rupees Origin : July, 2008

(*i*) Based on the past several years, monthly sales during January have been around Rs. 50,000. What is the typical seasonal relative for January?

(*ii*) If your seasonal relative for January was greater than 100, does it necessarily indicate that at least one of the 11 months has a seasonal relative, that is less than 100? Explain.

37. Fit a straight line trend to the following data and show the original observations and trend values on the graph paper :

Year	:	2004	2005	2006	2007	2008	2009	2010
Gross ex-factory value of output	:	672	824	967	1204	1464	1758	2057

[$Y = 1278 + 23.86X$]

38. The number of units of a product exported during 2003-2010 is given below. Fit a straight line trend to the data. Plot the given data showing also the trend.

Year	:	2003	2004	2005	2006	2007	2008	2009	2010
No. of units (in thousands)	:	12	13	13	16	19	23	21	23

[$Y = 17.5 + 0.893X$]

39. Calculate seasonal indices by the 'ratio-to-moving average method' from the following data :

Year	*I Quarter*	*II Quarter*	*III Quarter*	*IV Quarter*
2008	68	62	61	63
2009	65	58	66	61
2010	68	63	63	67

[105.3; 95.21; 100.97; 98.52]

40. The sales of a company increased from Rs. 4,00,000 in March to Rs. 4,80,000 in July 2010. The company's seasonal indices for these two months are 105 and 140 respectively. The owner of the company expressed dissatisfaction with the July sales but the Sales Manager said that he was quite pleased with the Rs. 80,000 increase. What argument should the owner of the company have used to reply the Sales Manager?

The Sales Manager also predicted on the basis of the July sales that the total 2012 sales were going to be Rs. 6,76,000. Criticise the Sales Manager's estimate.

41. The following table shows the number of salesmen working in a certain concern :

Year	:	2006	2007	2008	2009	2010
No. of salesmen :		28	38	46	40	56

Use the method of least squares to fit a straight line trend and estimate the number of salesmen in 2015.

42. The materials manager of a company has projected 10, 15 and 18 truckloads of a product for three consecutive months. The seasonal indices for these are 141.5, 125.8 and 82.6 respectively. Work out the seasonalised forecast for each month of three months.

43. The seasonal indices of the sale of readymade garments of a particular type in a departmental store are given below :

	Quarter	*Seasonal index*
I	Jan.—March	95
II	April—June	80
III	July—Sept.	90
IV	Oct.—Dec.	125

If the total sales in the first quarter of the year be worth Rs. 50,000, determine how much worth of garments of this type should be kept in store to meet the demand in each of the remaining quarters.

[50,000; 42,145.26; 47,368.42; 65,789.47]

44. A company estimates its sales for a particular year to be Rs. 24,00,000. The seasonal indices for sales are as follows :

Months	*Seasonal index*	*Months*	*Seasonal index*
January	75	July	102
February	80	August	104
March	98	September	100
April	128	October	102
May	137	November	82
June	119	December	73

Using the given information, calculate estimates of monthly sales of the company. Assume that there is no trend.

(MBA, Osmania Univ., 2002)

45. The following figures are the production data of a cement factory :

Year	*Production ('000 tonnes)*	*Year*	*Production ('000 tonnes)*
2000	17	2006	35
2001	20	2007	35
2002	19	2008	51
2003	26	2009	74
2004	24	2010	79
2005	40		

Fit the trend of the type $Y = a + bX + cX^2$ to the above data. Select the year 2005 as the working origin.

46. Using the data given below, explain how would you determine seasonal fluctuations in a time series :

Year	*Summer*	*Monsoon*	*Autumn*	*Winter*
2006	30	81	62	199
2007	33	104	86	171
2008	42	153	99	221
2009	56	172	129	235
2010	67	201	136	302

47. The number of units produced during 2003-2010 are given below :

Year :	2003	2004	2005	2006	2007	2008	2009	2010
Units produced :	56	55	51	47	42	38	35	32

(*i*) Fit a straight line trend and obtain the trend values.

(*ii*) Eliminate the trend. What components of the time series are thus left over?

(*iii*) What is the monthly increase in the number of units produced?

48. Compute a nonlinear trend of the form $Y = a + bX + cX^2$ for the data showing the production of wheat (in thousand tonnes) during the years 2002 to 2010.

Year	:	2002	2003	2004	2005	2006	2007	2008	2009	2010
Production of wheat ('000 tonnes)	:	9	10	12	15	13	10	8	16	25

(Take the year 2006 as working origin)

49. Find the trend values by the method of least squares for the following time series :

Year	:	2003	2004	2005	2006	2007	2008	2009	2010
Production ('000 tonnes)	:	351	366	362	400	419	420	450	518

Estimate the likely production for the year 2013.

50. Use method of least squares to determine sales for the year 2012.

Year	:	2006	2007	2008	2009	2010
Sales	:	100	110	130	125	160

51. Fit a straight line trend by the method of least squares to the following data :

Year	:	2000	2001	2002	2003	2004	2005	2006	2007
Earnings (Rs. Lakh)	:	38	40	65	72	69	60	87	95

(M.Com. Madurai-Kamaraj Univ., 2007)

52. The projected number of women of child bearing age (15–49) for India from 2000 to 2007 are as follows :

Year	:	2000	2001	2002	2003	2004	2005	2006	2007
No. of Women (in millions)	:	152.6	156.4	160.3	164.4	168.5	172.7	176.9	181.2

Fit a trend line. *(MBA, Anna Univ., 2007)*

53. What is meant by moving average ? Find the trend for the following series by three year weighted average with weights 1, 2, 1 :

Year (coded value)	:	–3	–2	–1	0	1	2	3
Sales (in thousand units)	:	2	4	5	7	8	10	13

(MBA, M.D. Univ., 2006)

54. The following are the annual profits in lakh of rupees, in a certain business :

Year	*Profits (Rs. lakh)*	*Year*	*Profits (Rs. lakh)*
2004	60	2008	80
2005	72	2009	85
2006	75	2010	95
2007	65		

(*i*) Use the method of least square to fit a straight line trend to the above data.

(*ii*) Also make an estimate of the profits for the year 2014.

* * * *

11

Probability

INTRODUCTION

The concept of probability which originated in the seventeenth century has become one of the most fascinating and debatable subjects in recent years. The probability formulae and techniques were developed by Jacob Bernoulli (1654-1705), De Moivre (1667-1754), Thomas Bayes (1702-1761) and Joseph Lagrange (1736-1813). In the nineteenth century, Pierre Simon, Laplace (1749-1827) unified all these early ideas and compiled the first general theory of probability. In fact, volumes have been written on probability and still the intellectual controversy concerning the foundations of probability theory is going on. So many people use the concept of probability in their daily lives without actually being aware of it. For example, we often find people making such statements as: 'It is likely that it may rain', 'We probably will get the contract', 'It is possible that the price of shares may go down further', etc. Though such assertions seem to be quite clear, a careful analysis would reveal that there are considerable difficulties in specifying the meaning of these statements.

In the beginning, the probability theory was successfully applied at the gambling tables. Gradually it was applied in the solution of social, economic, political and business problems. The insurance industry, which emerged in the 19th century, required precise knowledge about the risk of loss in order to calculate premiums. Within a few decades many learning centres were studying probability as a tool for understanding social phenomena. Today the concept of probability has assumed great importance and the mathematical theory of probability has become the basis for statistical applications in both social and decision-making research.

In fact, probability has become a part of our everyday lives. In personal and management decisions, we face uncertainty and use probability theory, whether or not we admit the use of something so sophisticated. We live in a world in which we are unable to forecast the future with complete certainty. Our need to cope with uncertainty leads us to the study and use of probability theory. In many instances, we, as concerned citizens, will have some knowledge about the possible outcomes of a decision. By organizing this information and considering it systematically, we will be able to recognise our assumptions, communicate our reasoning to others, and make a sounder decision than we could by using a shot-in-the-dark approach.

Probability constitutes the foundation of statistical theory and application. Knowledge of probabilistic methods has become increasingly essential in quantitative analysis of business and economic problems. In particular, probability theory is a basic component of the formal theory of decision-making under risk and uncertainty. Probability measures provide the decision-maker with the means for quantifying the uncertainties which affect his choice of appropriate actions. A through understanding of the fundamentals of probability theory will permit a businessman to deal with uncertainty in business situations in such a way that he can assess systematically the risks involved in each alternative, and consequently act to minimize risks.

Over the years, numerous definitions are given and we can classify these into different schools of thought. There are mainly four schools of thought on probability, namely:

1. The classical or *a priori* approach,
2. The relative frequency or empirical approach,
3. The axiomatic approach, and
4. The personalistic approach.

1. The Classical Approach

The classical or *a priori* approach happens to be the earliest. This school of thought assumes that all the possible outcomes of an experiment are mutually exclusive and equally likely. The words "equally likely" convey the notion of equally probable, and mutually exclusively means if one event occurs the other event will not occur, *i.e.*, the classicists believe that each outcome of an experiment has the same chance of appearing as any other and, therefore, can be assigned the same weight (probability) for its occurrence as any other. For example, when we toss a coin the probability of head is equal to the probability of a tail and is equal to 1/2. Similarly, each card drawn at random from well-shuffled deck of playing cards has the same chance to be drawn *i.e.*, 1 in 52 or 1/52, the probability of drawing a heart would be 13/52=1/4, the probability of drawing a black card 36/52=1/2, etc. Thus the classical concept defines the probability of an event as follows: If there are 'a' possible outcomes favourable to the occurrence of an event E, and 'b' possible outcomes unfavourable to the occurrence of E and all these possible outcomes are equally likely and mutually exclusive, then the probability that the event E will occur, denoted by $P(E)$, is

$$P(E) = \frac{a}{a+b} = \frac{\text{Number of outcomes favourable to the occurrence of event } E}{\text{Total number of outcomes}}$$

In the *a priori* method of measurement as well as in all other methods, the probability of an event E is a number such that $0 \leq P(E) \leq 1$, and the sum of the probability that an event will occur and the probability that it will not occur is equal to one.

The classical approach has two interesting characteristics: first, the subjects referred to as fair coins, fair deck of cards, true dice are abstractions in the sense that no real world object exactly possesses the features postulated. If a coin is unbalanced or there is a loaded die, the classical approach of assigning equal probability would offer us nothing but confusion. Secondly, in order to determine probabilities in the above examples, no coins had to be tossed, no cards shuffled, nor dice rolled, *i.e.*, no experimental data were required to be collected. The probability calculations were based entirely upon logical prior (thus, *a priori*) reasoning.

2. Relative Frequency Approach

While the classical theory is useful for solving problems which involve games of chance, it encounters serious difficulties in analysing a wide range of other types of problems. For example, it is inadequate for answering questions such as: What are the probabilities that (*a*) a man aged 45 will die within the next year, (*b*) a consumer in a certain metropolitan area will purchase a particular product during the next month, (*c*) a production process used by a particular firm will produce a defective item.

In none of these situations, it is feasible to establish a set of complete and mutually exclusive outcomes each of which is equally likely to occur. For example, in (*a*) there are only two possible occurrences, the individual will die during the ensuing year or he will live. The likelihood that he will die is of course, much smaller than he will live. How much smaller? This is the type of question that requires reference to empirical data.

The relative frequency theoreticians agree that the only valid procedure for determining event probabilities is through repetitive experiments. For example, when a coin is tossed, what is the probability that the coin will turn up heads? The relative frequency theorist would actually toss the coin and calculate the proportion of times our coin falls heads. Suppose he tosses the coin 50 times and it falls head 20 times, then the ratio 20/50 is used as the estimate of the probability of heads of this coin. It may be noted that even if the coin is perfect, one may not get exactly 25 heads out of tosses. In other words, it is not possible to obtain the true probability from repeated experiments. However, if the coin were perfect, the estimate would approach the true ratio (probability) as the number of trails increased. Thus, if the coin is tossed 200 times, we may have 85 heads (or 115 tails). The relative frequency becomes 85/200 = 0.425. If we further toss the coin 2,000 times, we may have 980 heads (or 1020 tails); the relative frequency being 980/2000 = 0.49, and so on. Since the probability of an event is determined objectively by repetitive empirical observations, the relative frequency theory is also called the objective or empirical definition of probability.

The ratio of the number of occurrences of an event to the number of possible occurrences in an experiment is referred to as the relative frequency. Two definitions of probability in terms of relative frequency can be given:

(*a*) If an experiment is performed n times under the same conditions and there are 'a' outcomes, $a \leq n$, favouring an event, then an estimate of the probability of that event is the ratio a/n.

(*b*) The estimate of probability of event, a/n approaches a limit, the true probability of the event, when n approaches infinity is given by

$$P(E) = \underset{n \to \infty}{Limit} \frac{a}{n}$$

It may be noted that we can never obtain the probability of an event as given by the above limit. In practice, we can only try to have a close estimate of $P(E)$ based on large n. However, this approach does emphasise that probability involves a long-run concept.

3. The Axiomatic Approach*

The classical approach restricts the calculation of probability to essentially equally likely and mutually exclusively events. The resolution of non-mutually exclusive events of reality into mutually exclusive subevents and the introduction of 'equal likelihood' among events which are essentially not so in reality are questions not clearly treated by the classicists. On the other hand, the empirical or relative frequency approach requires that every question of probabilistic nature be examined experimentally in the laboratory of the mathematician under identical conditions, and that too over a very long period of time, through the process of repeated observations, if estimates of the chances of occurrence of the events under consideration are required.

The axiomatic theory of probability is an honest attempt at constructing a theory of probability, largely free from the inadequacies of both the classical and empirical approaches, in the true mathematical tradition. It is true that the introduction of advanced logic through mathematical abstractions renders the complex real-world situation too idealised (or too simplified) to be of any immediate practical utility. But nonetheless it plays an important role in rendering a reasonable amount of comprehensibility and tractability to the understanding of myriad chance phenomena observed in nature, at least in the initial stages of any scientific inquiry into their structure and composition, where other approaches have at best left them less comprehensible and less tractable. Thus, the primary purpose of the development of an

*For details, see *Probability Theory* by M, Loeve Van Nostrand.

axiomatic theory lies in the fact that it makes available to the inquisitive mind a large body of abstract mathematical concepts, tools and techniques with which to identify, model, study and infer about real-world chance phenomena of interest. For a reasonably complete description of reality will not be 'complete' unless some amount of 'abstraction was made somewhere along the course of the inquiry, and models are nothing but abstractions of reality in some necessary degree.

4. The Personalistic Approach†

Though this approach to probability is relatively recent its application to statistical problems has occurred virtually entirely in the post-world World War II Period, particularly in connection with statistical decision theory. According to the personalistic or subjective concept, the probability of an event is the degree of confidence (or belief) placed in the occurrence of an event by a particular individual based on the evidence available to him. This evidence may consist of relative frequency to data and any other quantitative or qualitative information. According to the degree of belief for its possible occurrence, a subjectivist would assign a weight between 0 and 1 to an event. Thus, if one believes that it is very likely that the event will occur, he will assign it a probability close to one and if he believes that it is unlikely that the event will occur, he will assign a probability close to zero.

This approach is very broad and flexible one, permitting probability assignments to events for which there may be no objective data, or for which there may be a combination of objective and subjective data. The subjective approach grants that different reasonable individuals may differ in their degree of confidence even when offered the same evidence and consequently personal probabilities for the same event may differ in the eyes of different decision-makers.

Though broadly there are four different schools of thought on probability, there is hardly any disagreement on the foundation of probability at the mathematical level as each school defines probability as a ratio or proportion. Each viewpoint has its own merits and depending upon the problem under consideration one may use whichever approach is appropriate and convenient.

Elements of Set Theory

Modern approach to probability theory generally employs set theory and it will be used here for the development of some fundamental concepts and tools.

Sets. A set is any well-defined specified collection of distinct elements or objects. The objects which comprise the set are usually referred to as elements or members of the set and are said to belong to that set or to be contained in it. The set must be well specified or well defined in the sense that it must be possible at least in principle, to specify the set so that one can decide whether any given member does or does not belong to the set. The members of the set are distinct in the sense that repetition of elements is not permitted in specifying the set. The collection of aggregation or totality of elements is referred to simply as a set denoted by S. Thus the following collections are examples of sets :

The students enrolled in a university.

The books in a departmental library.

The employees of a company.

The citizens of India.

The possible outcomes of the roll of a single die.

A set is usualy described in either of the following two ways:

A roster or tabulation method, and the rule or defining property method.

---

†The concept was first introduced in 1926 by Frank Ramsey who presented a formal theory of personal probability in his book: *The Foundation of Mathematics and Other Logical Essays* (London: Kegan Paul; New York: Harcourt Brace and World 1931).

Roster or Tabulation Method

The elements are usually enclosed within brackets. For example, the set consisting of the possible outcomes (Tail = T, Head = H) of single toss of a coin may be expressed as:

$$S = \{T, H\}$$

The set of possible outcomes of tossing two coins may be written as:

$$S = \{T, T), (T, H), (H, T), (H, H)\}$$

The order in which the elements of a set are listed is of no importance. It is important, however, that each element be listed only once. Note that in the second example there are four elements in the set, *viz.*, (T, T), (T, H), (H, T) and (H, H).

Rule or Defining Property Method

Sometimes it is helpful to have a brief and exact way to describe sets without listing elements. For example, the set of all university students may be expressed as :

$$S = \{x/x \text{ is a student in the university}\}$$

We read this as "S is the set of all x such that x is a student in the university."

Universal set

The universal set U is defined as that set consisting of all the elements under consideration. Thus if A is any set and U is the universal set, then every element in A must be in U (since it consists of elements under consideration).

Null Set

A set having no element at all is called a null or an empty set. The symbol used to denote it is a Greek letter ø (Phi).

Subset

If every element of a set A is also an element of a set B, then A is called a subset of B. For example, consider the set $A = (3, 5)$ and the set $B = (1, 2, 3, 4, 5)$. We note that every element in the set A is also an element of the set B. The set A is said to be the subset of B. Symbolically, we write this as $A \subset B$ read as A is contained in B or A is a subset of B.

Equal Sets

Two sets A and B are said to be equal if and only if every element of A is also an element of B and *vice versa*. Symbolically, $A = B$ if and ony if $A \subset B$ and $B \subset A$.

Set Operations

We shall now consider certain operations on sets that will result in the formation of new sets.

Intersection of Sets

The intersection of two sets A and B is the set of elements that are common to both A and B. Symbolically, the intersection of A and B is written as $A \cap B = \{x / x \in A \text{ and } x \in B\}$. In the following diagram, the shaded area corresponds to the intersection of sets A and B. U is the universal set.

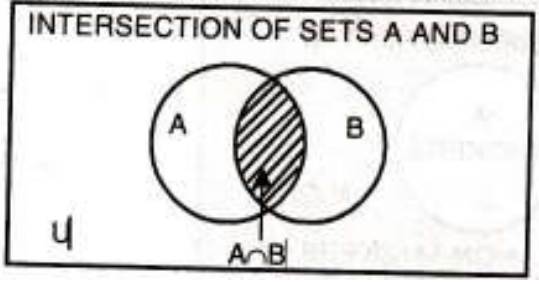


Illustration 1. Consider the sets of numbers :
$U = \{x/x \text{ is positive integer}\}$
$A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
$B = \{8, 9, 10, 11, 12, 13, 14\}$
Then $A \cap B = \{8, 9, 10\}$.
Since only these elements appear in both A and B.

Disjoint sets

Two sets A and B are called disjoint if they do not intersect. This can be expressed as $A \cap B = \varnothing$ where ø is a null set. When the two sets do not intersect, they are said to be disjoint or mutually exclusive. These sets are shown below in the diagram.

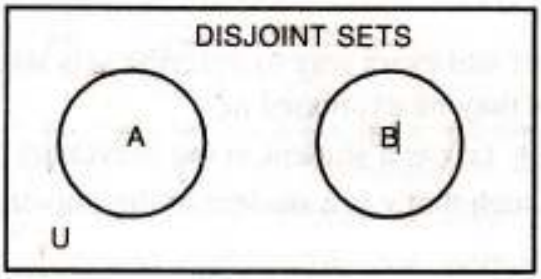


Illustration 2. Consider the sets of numbers :
$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
$A = \{1, 2, 3\}$
$B = \{5, 6\}$
Note that A and B do not intersect. This can be expressed as $A \cap B = \phi$.

Union of sets

The union of two sets A and B is the set of elements that belong either to A or B or both. this is expressed as $A \cup B = \{x / x \in A \text{ or } x \in B\}$. The union of two sets sometimes is expressed as the logical sum of the two sets. In the following diagram, the area representing the elements of the set $A \cup B$ has been shaded.

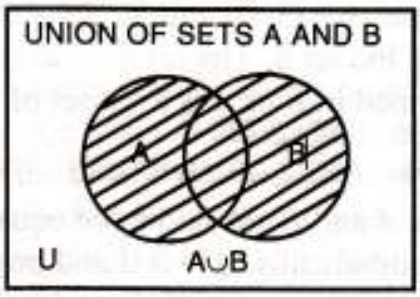
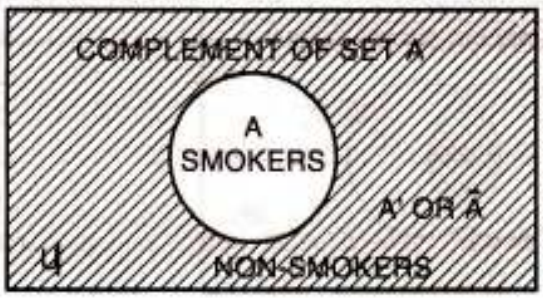


Illustration 3. Consider the set of numbers :
$U = \{x/x \text{ is a positive integer}\}$
$A = \{1, 3, 5\}$
$B = \{3, 4, 5, 6\}$
Then $A \cup B = \{1, 3, 4, 5, 6\}$
Since these elements appear in either A or B or both.

Complement of a Set

If A is a subset of the universal set U, then the complement of A with respect to U is the set of all elements of U that are not in A or the complement of set A is the set of all elements that do not belong to A and is denoted by A' or $\overline{A}$. In symbols, $A' = [x/x \in U \text{ and } x \notin A]$. Suppose we consider the employees of a firm as the universal set. Let all the smokers form a subset. Then all the non-smokers also form a subset which is called the complement of the set constituting smokers. In the following diagram, the area representing the complement of A has been shaded.



Here A and A' do not intersect—that is, $A \cap A' = \phi$. Hence A and A' are mutually exclusive. Another characteristic of this case is that $A \cup A' = U$. Thus A and A' are also completely exhaustive.

Illustration 4. Let $U = \{1, 2, 3, 4\}$
$A = \{1, 2, 3\}$
then $A' = \{4\}$

Difference of Two Sets

The difference of sets A and B is defined as $A-B = (x/x \in A$ and $x \notin B)$. This is shown below as the shaded area.

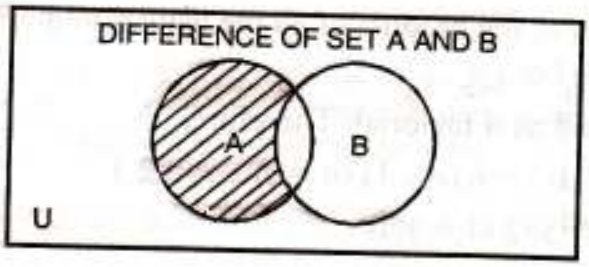


Illustration 5. Let $A = \{1, 3, 5, 7, 9, 11, 13\}$
$B = \{5, 9, 13, 17\}$
then $A - B = \{1, 3, 7, 11\}$

The following illustration will explain the use of different set operations :

Illustration 6. A firm has 231 employees classified by age and job category as follows :

Job	Category	Age Category					Total
		A_1	A_2	A_3	A_4	A_5	
		≤ 20	21–25	26–30	31–35	>35	
B_1	Peons	20	20	15	10	5	70
B_2	Clerks	3	6	3	2	1	15
B_3	Draftsmen	15	30	35	20	10	110
B_4	Salesmen	1	5	10	5	2	23
B_5	Junior Executives	0	1	5	2	0	8
B_6	Executives	0	0	2	2	1	5
	Total	39	62	70	41	19	231

Based on the above table, explain in words the following sets and give the number of employees in each set:

(*a*) $B_1 \cap A_5$
(*b*) $A_2 \cap B_6$
(*c*) $B_4 \cap A_5$
(*d*) $A_1 \cup B_6$
(*e*) $A_3 \cup A_5$
(*f*) $B_2 \cup B_3$
(*g*) A_4'
(*h*) $(A_1 \cup A_2) \cap B_3$
(*i*) $(B_3 \cup B_4) \cap A_5$

Solution. (*a*) $B_1 \cap A_5$ gives us the intersection between peons and age greater than 35, *i.e.*, those peons who are more than the age 35. From the table this gives us the value 5. Hence $B_1 \cap A_5 = 5$.

(*b*) Similarly, $A_2 \cap B_6$ means the intersection between the age group 21–25 and that of executives, *i.e.*, the executives who are in the age group 21–25. From the table this value is 0. Hence $A_2 \cap B_6 = 0$.

(*c*) Similarly, $B_4 \cap A_5 = 2$.

(*d*) $A_1 \cup B_6$ gives the union between age less than 20 and executives, *i.e.*, either in the category of age less than 20 and executives or both, therefore, $A_1 \cup B_6 = 39 + 5 = 44$.

(*e*) Similarly, $A_3 \cup A_5 = 70 + 19 = 89$.

(*f*) Similarly, $B_2 \cup B_3 = 15 + 110 = 125$.

(*g*) A'_4 means not contained in A_4, *i.e.*, all those except the set A_4 containing 41 employees. Hence $A_4 = 231 - 41 = 190$.

(*h*) $(A_1 \cup A_2) \cap B_3$ gives us the union of A_1 and A_2 first then its intersection with B_3, *i.e.*, $(A_1 \cup A_2) \cap B_3 = 15 + 30 = 45$.

(*i*) Similarly, $(B_2 \cup B_4) \cap A_5 = 10 + 2 = 12$.

Counting Techniques

In computing the probability of an event, or the probability of a combination of events, when the total number of possible events is large, it will be convenient to have available some methods for counting the number of such events. In this section, some techniques to facilitate the counting of events will be presented. These are useful for counting number of events comprising the numerator and/or the denominator of a probability.

Factorials

Given the postive integer n, the product of all the natural numbers from n down through 1 is called n factorial and is written as n ! or $\lfloor n$.

The expression n ! is read as n factorial. Therefore,

$$n ! = n (n - 1) (n - 2) \ldots\ldots 3.2.1. \qquad \ldots(i)$$

For example : $5 ! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

From (i), we get $n ! = n (n\text{-}1) !$ and $0 ! = 1$*

Permutations

A permutation of a number of objects is an arrangement of these objects in a definite order. The number of permutations of a set of n objects, taken all together is n !. We have, in permutation, n spaces to fill. The first space can be filled with any one of the n objects and so in n ways. After this has been done (in any of the n ways), the second space can be filled with any of the remaining $(n - 1)$ objects; likewise, the third place can be filled in $(n - 2)$ ways, the fourth space in $(n - 3)$ ways and so forth. Therefore, the number of ways of filling n spaces is

$$n (n - 1) (n - 2) \ldots\ldots 3.2.1.$$

which is nothing but n !.

Denoting this by ${}^{n}P_{n}$, we have

${}^{n}P_{n} = n !$

Illustration 7. There are four clerks in an office whose tables are arranged in a line against a wall. How many seating arrangements are possible if each clerk can sit at any table?

Solution. $4 ! = 4.3.2.1 = 24$ ways.

The total number of arrangements of n objects taken r at a time with $r \leq n$ denoted as ${}^{n}P_{r}$, is

$${}^{n}P_{r} = \frac{n!}{(n-r)!}$$

The permutation of n objects taken r at a time can also be denoted by the following symbols

$$P(n,r),\ {}_{n}P_{r},\ P_{n,r},\ P^{n}{}_{r}$$

Illustration 8. A personnel manager has received requisitions for one typist each from the Production department, Marketing department and Research department. There are seven applicants available from which these three positions may be filled. In how many ways three typists be selected from the seven applicants and assigned to the three different openings?

Solution. There are seven ways to fill the first position after which there are six ways to fill the second position after which there are five ways to fill the third position. This gives $7 \times 6 \times 5 = 210$ ways. The same result can be obtained using the formula

$${}^{7}P_{3} = \frac{7!}{(7-3)!} = \frac{7 \times 6 \times 5 \times 4!}{4!} = 7 \times 6 \times 5 = 210.$$

---

*Note that $n! = n (n - 1) !$ or $(n - 1) ! = \frac{n!}{n}$

Letting $n = 1$, $0! = 1$.

Combinations

A combination of number of objects is a selection of these objects, considered without regard to their order. The total number of combinations of a set of n objects taken r at a time ($r \leq n$), usually denoted by ${}^nC_r$ and is given by

$$ {}^nC_r = \frac{n!}{r!(n-r)!} $$

This can also be denoted by the following symbols.

$$ \binom{n}{r},\ {}_nC_r,\ C_{n,r},\ C^n{}_r $$

Illustration 9. A sales manager has seven field representatives working under him. A local consulting firm at a fee of Rs. 1500 per man, is conducting a three-day seminar on sales to which the sales manager would like to send all the seven of his field representatives. However, his budget will allow him to send only three men. How many different ways are there for him to compose this group of three men?

Solution. The number of possible combinations of three men selected from a set of seven men is

$$ {}^7C_3 = \frac{7!}{4!3!} = \frac{7.6.5.4!}{4!3.2.1} = 35. $$

Note: *It is important to note that in a permutation, order counts; in a combination, order does not count.*

Illustration 10. *Out of 5 mathematicians and 7 statisticians, a committee consisting of 2 mathematicians and 3 statisticians is to be formed. In how many ways can this be done if (a) any mathematician and any statistician can be included (b) one particular statistician must be on the committee, (c) particular mathematician cannot be on the committee?*

Solution. (*a*) 2 mathematicians out of 5 can be selected in ${}^5C_2$ ways.

3 statisticians out of 7 can be selected in ${}^7C_3$ ways.

Total number of possible selection $= {}^5C_2 \times {}^7C_3 = 10 \times 35 = 350$

(*b*) 2 mathematicians out of 5 can be selected in ${}^5C_2$ ways.

2 additional statisticians out of 6 can be selected in ${}^6C_2$ ways.

Total number of possible selections $= {}^5C_2 \times {}^6C_2 = 10 \times 15 = 150.$

(*c*) 2 mathematicians out of 4 can be selected in ${}^4C_2$ ways.

3 statisticians out of 7 can be selected in ${}^7C_3$ ways.

Total number of possible selections $= {}^4C_2 \times {}^7C_3 = 3 \times 35 = 105.$

Random Experiment

A random experiment is a well-defined process of observing a given chance phenomena through a series of trials (finite or infinite) each of which leads to a single outcome.

Observation of chance phenomena is called random experiment so as to distinguish them from experiments under control conditions, for example in a physical laboratory.

Events

And event is a possible outcome of an experiment or a result of a trial or an observation.

Elementary Events

An elementary event or a simple event is a single possible outcome of an experiment. It is thus an event which cannot be further subdivided into a combination of other events.

Compound Events

When two or more events occur in connection with each other, then their simultaneous occurrence is called a compound event. The compound event is an aggregate of simple events.

Mutually Exclusive Events

Two events are said to be mutually exclusive when both cannot happen simultaneously in a single trial or, in other words, the happening of one prevents the happening of the other and *vice versa.* For

example, if a single coin is tossed either head can be up or tail can be up, both cannot be up at the same time. Similarly, a person may be either alive or dead at a certain time, he cannot be both alive as well as dead at the same time.

Collectively Exhaustive Events

In the example of fair coin tossing, there are two possible outcomes: head and tail. The list of these outcomes is collectively exhaustive since the result of any toss must be either head or tail. Collective exhaustive events are those which include all possible outcomes. The sum of the probabilities must be one for mutually exclusive and collectively exhaustive events.

Complementary Events

Let A be an event of the number of favourable cases in the experiment, then $\overline{A}$ called the complementary event of A is the number of nonfavourable cases in the experiment. Clearly the events A and $\overline{A}$ are mutually exclusive and collectively exhaustive.

Equally likely Events

Events are said to be equally likely when one does not occur more often than the others. For example, if an unbiased coin or die is thrown, each face may be expected to be observed approximately the same number of times in the long run. Similarly, the cards of a pack of playing cards are so closely alike that we expect each card to appear equally often when a large number of draws are made with replacement. Random number tables are based on this concept.

PROBABILITY LAWS

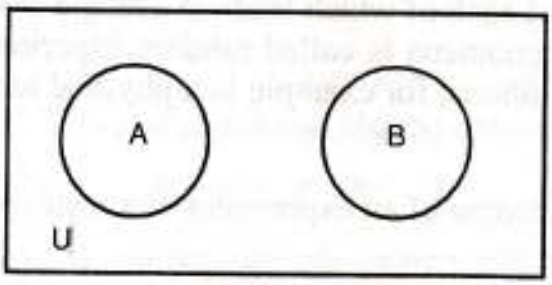
There are several laws that can ease our task of computing probabilities. In this section, we shall discuss two of the fundamental laws of computing probabilities, *viz.*, Addition law and Multiplication law.

Addition Law

The probability of occurrence of either event A or event B of two mutually exclusive (or disjoint sets) events is equal to the sum of their individual probabilities. Symbolically, we may write,

$$P(A \cup B) = P(A) + P(B)$$

DISJOINT EVENTS



Since A and B can be written as a union of simple events in which no simple event of B appears in A, hence, the result follows.

If two events A and B are not mutually exclusive (joint events) then the addition law can be stated as follows:

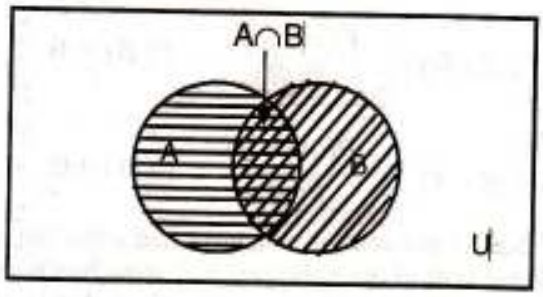
The probability of the occurrence of either event A or event B or both is equal to the probability that event A occurs, plus the probability that event B occurs minus the probability that both events occur. Symbolically, it can be written as

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

Proof. $$P[A \cup B] = \frac{n(A \cup B)}{n(U)}$$

where $n(A \cup B)$ indicates the number of elements belonging to $A \cup B$, and $n(U)$ is the total number of elements in the universal set U.

JOINT EVENTS



$$P(A \cup b) = \frac{n(A) + n(B) - n(A \cap B)}{n(U)}$$

[By adding $n(A)$ and $n(B)$, we count twice $(A \cap B)$. See diagram above.]

$$= \frac{n(A)}{n(U)} + \frac{n(B)}{n(U)} - \frac{n(A \cap B)}{n(U)}$$

$$= P(A) + P(B) - P(A \cap B).$$

Illustration 11. City residents were surveyed recently to determine readership of newspapers available. 50% of the residents read the morning paper, 60% read the evening paper, and 20% read both newspapers. Find the probability that a resident selected reads either the morning or evening paper or both the papers.

Solution. Let A and B represent the events that the resident read morning and evening paper respectively.

Then $$P(A) = 0.50;\ P(B) = 0.60;\ \text{and}\ P(A \cap B) = 0.20$$

The probability that the resident reads either the morning or evening or both the papers is given by :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= 0.50 + 0.60 - 0.20 = 0.90.$$

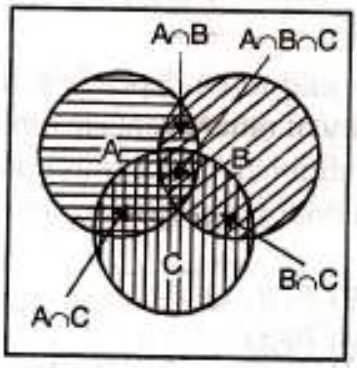
Generalisation

The addition law for mutually exclusive events can be extended to cover any number of events. In particular, if A, B and C are three mutually exclusive events, then the probability that any one of these events will occur is given by

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

When the events are not mutually exclusive, then the formula becomes

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$
$$- P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$



Conditional Probability

When we are dealing with probabilities of a subset rather than of the whole set, our attention is focused on the probability of an event in a subset of the whole set. Probabilities associated with the events defined on the subsets are called conditional probabilities. The conditional probability of A, given B, is equal to the probability of $A \cap B$ divided by the probability of B, provided that the probability of B is not zero. Symbolically, we may write this as

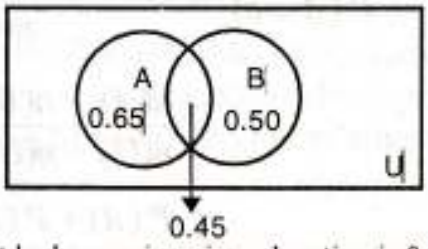
$$P(A/B) = \frac{P(A \cap B)}{P(B)} \; ; \; P(B) \neq 0.$$

Similarly $$P(B/A) = \frac{P(A \cap B)}{P(A)} \; ; \; P(A) \neq 0.$$

Illustration 12. A study showed that 65 per cent of managers had some business education and 50 per cent had some engineering education. Furthermore, 20 per cent of the managers had some business education but no engineering education. What is the probability that a manager has some business education, given that he has some engineering education?

Solution. Let A denote the event that the manager has some business education and B denote that he has some engineering education.

Then $$P(A) = 0.65, \; P(B) = 0.50, \; P(A \cap B) = 0.45$$



Therefore $$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.45}{0.50} = 0.9$$

Hence the required probability that a manager has some business education given that he has engineering education is 0.9.

Multiplication Law

The multiplication law may be stated as follows :

The probability of the joint occurrence of event A and event B is equal to the conditional probability of A given B, times the probability of B.

Symbolically, we write

$$P(A \cap B) = P(A/B) \times P(B)$$

or $$P(B \cap A) = P(B/A) \times P(A)$$

Proof.

$$P(A/B) = \frac{n(A \cap B)}{n(B)} = \frac{n(A \cap B)/n(U)}{n(B)/n(U)} = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A/B) \; . \; P(B).$$

Generalisation. The multiplication law can be extended for more than two events. If we have three events A, B and C which are not mutually exclusive then the formula becomes

$$P(A \cap B \cap C) = P(A) \; . \; P(B/A) \; . \; P(C/A \cap B)$$

For n events $A_1, A_2, \ldots\ldots, A_n$, the formula becomes

$$P(A_1 \cap A_2 \cap A_3 \cap \ldots\ldots \cap A_n) = P(A_1) \; P(A_2/A_1) \; P(A_3/A_1 \cap A_2 \ldots\ldots)$$

$$\ldots\ldots\ldots P(A_n/A_1 \cap A_2 \cap \ldots\ldots \cap A_{n-1})$$

Dependent Events. Two events are said to be dependent if the occurrence or non-occurrence of one event in any trial affects the probability of other events in other trials. Thus, in the case of dependent events, the probability of any event is conditional, or depends upon the occurrence or non-occurrence of other events. From definitions of conditional probabilities, we can see that if A and B are dependent events,

$$P(A \cap B) = P(A/B) \; P(B) \qquad \ldots(i)$$

or $$P(B \cap A) = P(B/A) \; P(A) \qquad \ldots(ii)$$

The order is of no significance in the intersection of two events, since $A \cap B = B \cap A$. Therefore, we get an important property of intersection, *viz.*,

$$P(A \cap B) = P(A/B) \cdot P(B) = P(B/A) \cdot P(A)$$

Independent Events. Two events are said to be independent, if the probability of the occurrence of one event will not affect the probability of the occurrence of the second event. Independent events are those events whose probabilities are in no way affected by the occurrence of any other event preceding, following or occurring at the same time.

Two events A and B are said to be independent if and only if

$$P(A \cap B) = P(A)\, P(B)$$

which implies from (*i*) and (*ii*), that

$$P(A/B) = P(A)$$

and

$$P(B/A) = P(B)$$

Illustration 13. A candidate is selected for interview of management trainees for 3 companies. For the first company there are 12 candidates, for the second there are 15 candidates and for the third there are 10 candidates. What are the chances of his getting job at least at one of the company?

Solution. The probability that the candidate gets the job at least at one company = 1– probability that the candidates does not get the job in any company.

Probability that the candidate does not get the job in the first company

$$= 1 - \frac{1}{12} = \frac{11}{12}$$

Probability that the candidate does not get the job in the second company

$$= 1 - \frac{1}{15} = \frac{14}{15}$$

Probability that the candidate does not get the job in the third company

$$= 1 - \frac{1}{10} = \frac{9}{10}$$

Since the events are independent, therefore, the probability that the candidate does not get any job in any of the three companies

$$= \frac{11}{12} \times \frac{14}{15} \times \frac{9}{10} = \frac{231}{300} = 0.77$$

Hence the required probability = 1– 0.77 = 0.23.

Bayes' Theorem

It is associated with the name of Thomas Bayes (1702-1761) and is a theorem on probability, concerned with a method of estimating the probabilities of the causes by which an observed event may have been produced. This theorem may be stated as follows :

Let $B_1, B_2, \ldots\ldots, B_n$, be n mutually exclusive events whose union is the universe, and let A be an arbitrary event in the universe, such that $P(A) \neq 0$. Given that $P(A/B_i)$, and $P(B_i)$ $(i = 1, \ldots\ldots, n)$ are known.

$$P(B_j / A) = \frac{P(A/B_j)\, P(B_j)}{\sum_{i=1}^{n} P(B_j)\, P(A/B_j)} \quad \text{for } j = 1, \ldots\ldots, n.$$

This equation is called the formula for the probability of 'Causes', since it enables one to find the probability of a particular B_j, or 'Cause' by which the event A may have been brought about. It is sometimes written in another form as follows :

$$P(B_j / A) = \frac{P(A \cap B_j)}{P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)}$$

The Bayes' theorem is frequently used as a mechanism for revising the probability of an event after observing information about a process. The initial and revised probabilities are referred to as prior and posterior probabilities respectively.

Illustration 14. In a post office, three clerks are assigned to process incoming mail. The first clerk, B_1, processes 40 per cent, the second clerk, B_2, processes 35 per cent and the third clerk, B_3, processes 25 per cent of the mail. The first clerk has an error rate of 0.04, the second has an error rate of 0.06 and the third has an error rate of 0.03. A mail selected at random from a day's output is found to have an error. The Post Master wishes to know the probability that the mail was processed by the first, second, or third clerk, respectively.

Solution. Let A denote the event that a mail containing an error is selected at random and B_1, B_2 and B_3 be the event that the mail was processed by the first, second and third clerk respectively. Using our usual notation, we want to compute the conditional probabilities :

$$P(B_1|A),\ P(B_2|A),\ P(B_3|A)$$

From the information given, we have

$$P(B_1) = 0.40,\ P(B_2) = 0.35 \text{ and } P(B_3) = 0.25.$$

These probabilities, which can be obtained without additional information are called prior probabilities.

We are also given the information that the conditional probabilities observing a record with an error, given that it was processed by one of the three clerks are :

$P(A/B_1) = 0.04,\ P(A/B_2) = 0.06$ and $P(A/B_3) = 0.03$.

From these probabilities, we can calculate joint probabilities :

$P(A \cap B_1) = P(A/B_1)\ P(B_1) = 0.04 \times 0.40 = 0.016$

$P(A \cap B_2) = P(A/B_2)\ P(B_2) = 0.06 \times 0.35 = 0.021$

$P(A \cap B_3) = P(A/B_3)\ P(B_3) = 0.03 \times 0.25 = 0.0075$

Use Bayes' formula to obtain the desired probabilities.

$$P(B_1/A) = \frac{P(A \cap B_1)}{P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)}$$

$$= \frac{0.016}{(0.016 + 0.021 + 0.0075)} = \frac{0.016}{0.0445} = 0.36$$

Similarly,

$$P(B_2/A) = \frac{P(A \cap B_2)}{0.0445} = \frac{0.021}{0.0445} = 0.47.$$

$$P(B_3/A) = \frac{P(A \cap B_3)}{0.0445} = \frac{0.0075}{0.0445} = 0.17.$$

These probabilities are called posterior probabilities because they were calculated after it was known that the mail was one containing an error.

MISCELLANEOUS ILLUSTRATIONS

Illustration 15. The probability that a contractor will get a plumbing contract is 2/3 and the probability that he will not get an electric contract is 5/9. If the probability of getting at least one contract is 4/5, what is the probability that he will get both?

Solution. Let A and B denote the event that the contractor will get a plumbing and electric contract respectively.

Therefore $$P(A) = \frac{2}{3};\quad P(B) = 1 - \frac{5}{9} = \frac{4}{9};\quad P(A \cup B) = \frac{4}{5}$$

Hence $$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

$$= \frac{2}{3} + \frac{4}{9} - \frac{4}{5} = \frac{14}{45} = 0.31.$$

The required probability that the contractor will get both the contracts is given by 0.31.

Illustration 16. The probability that a management trainee will remain with a company is 0.60. The probability that an employee earns more than Rs. 50,000 per month is 0.50. The probability that an employee is a management trainee who remained with the company or who earns more than Rs. 50,000 per month is 0.70. What is the probability that an employee earns more than Rs. 50,000 per month, given that he is a management trainee who stayed with the company?

Solution. Let A = An employee who earns more than Rs. 50,000 per month.

B = A management trainee who stayed with the company.

Then $P(A) = 0.50;\ P(B) = 0.60;\ P(A \cup B) = 0.70$

To get the value of $P(A \cap B)$, we can use the following formula

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Substituting these values, we have

$$0.70 = 0.50 + 0.60 - P(A \cap B) \text{ or } P(A \cap B) = 0.40$$

Therefore, $$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.40}{0.60} = \frac{2}{3} = 0.67.$$

Hence the probability that an employee earns more than Rs. 50,000 per month given he is a management trainee is 0.67.

Illustration 17. The Human Resource department of a company has records which show the following analysis of the 200 engineers.

Age	*Bachelor's degree only*	*Master's degree*	*Total*
Under 30	90	10	100
30 to 40	20	30	50
Over 40	40	10	50
Total	150	50	200

If one engineer is selected at random from the company, find :

(*a*) the probability he has only a bachelor's degree.

(*b*) the probability he has a master's degree, given that he is over 40.

(*c*) the probability he is under 30, given that he has only a bachelor's degree.

Solution. Let us define the events A, B, C and D as follows :

A : An engineer is under 30 years of age.

B : An engineer is over 40 years of age.

C : An engineer has bachelor's degree only.

D : An engineer has a master's degree.

(*a*) The probability of an engineer who has a bachelor's degree only is

$$P(C) = \frac{150}{200} = 0.75.$$

(*b*) The probability of an engineer who has a master's degree, given that he is over 40 years is

$$P(D/B) = \frac{P(D \cap B)}{P(B)} = \frac{10/200}{50/200} = \frac{10}{50} = 0.20$$

(*c*) The probability of an engineer who is under 30 years, given he has only a bachelor's degree is

$$P(A/C) = \frac{P(A \cap C)}{P(C)} = \frac{90/200}{150/200} = \frac{90}{150} = 0.60$$

Illustration 18. An *MBA* applies for a job in two firms X and Y. The probability of his being selected in from X is 0.7 and being rejected at Y is 0.5. The probability of at least one of his applications being rejected is 0.6. What is the probability that he will be selected in one of the firms?

Solution. $$P(A) = 0.7;\ P(\overline{A}) = 1 - 0.7 = 0.3$$

$$P(B) = 0.5;\ P(\overline{B}) = 1 - 0.5 = 0.5;\ P(\overline{A} \cup \overline{B}) = 0.6$$

The probability that he will be selected in one of the firms is obtained by using addition rule.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

But $$P(A \cap B) = 1 - P(\overline{A} \cup \overline{B}) = 1 - 0.6 = 0.4$$

Hence $$P(A \cup B) = 0.7 + 0.5 - 0.4 = 0.8.$$

The probability of his being selected in one of the firms is 0.8.

Illustration 19. A problem in business statistics is given to five students : A, B, C, D and E. Their chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$ and $\frac{1}{6}$. What is the probability that the problem will be solved?

(MBA, Kumaon Univ., 2000)

Solution. Let E_1, E_2, E_3, E_4 and E_5 denote the events that the problem is solved by A, B, C, D and E respectively. Then we have :

$$P(E_1) = \frac{1}{2}; \qquad P(\bar{E}_1) = 1 - P(E_1) = \frac{1}{2}$$

$$P(E_2) = \frac{1}{3}; \qquad P(\bar{E}_2) = 1 - P(E_2) = \frac{2}{3}$$

$$P(E_3) = \frac{1}{4}; \qquad P(\bar{E}_3) = 1 - \frac{1}{4} = \frac{3}{4}$$

$$P(E_4) = \frac{1}{5}; \qquad P(\bar{E}_4) = 1 - \frac{1}{5} = \frac{4}{5}$$

$$P(E_5) = \frac{1}{6}; \qquad P(\bar{E}_5) = 1 - \frac{1}{6} = \frac{5}{6}$$

Since E_1, E_2, E_3, E_4 and E_5 are independent, the probability that all the five students fail to solve the problem is given by

$$P(\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_3 \cap \bar{E}_4 \cap \bar{E}_5) = P(\bar{E}_1)\, P(\bar{E}_2)\, P(\bar{E}_3)\, P(\bar{E}_4)\, P(\bar{E}_5)$$

$$= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdot \frac{4}{5} \cdot \frac{5}{6} = \frac{1}{6}.$$

Therefore, the required probability that the problem is solved

$$= 1 - \frac{1}{6} = \frac{5}{6} = 0.83.$$

Illustration 20. The odds that A speaks the truth is 3 : 2 and the odds that B speaks the truth is 5 : 3. In what percentage of cases are they likely to contradict each other on an identical point? *(MBA, DU, 1999)*

Solution. The probability that A speaks the truth = $P(A) = 3/5$

The probability that B speaks the truth = $P(B) = 5/8$

The probability that A tells a lie = $P(\bar{A}) = 2/5$

The probability that B tells a lie = $P(\bar{B}) = 3/8$

The probability that A speaks the truth and B tells a lie is given by

$$P(A)\, P(\bar{B}) = \frac{3}{5} \times \frac{3}{8} = \frac{9}{40}$$

The probability that B speaks the truth and A tells a lie is given by :

$$P(B)\, P(\bar{A}) = \frac{5}{8} \times \frac{2}{5} = \frac{10}{40}$$

The required probability = $P(A)\, P(\bar{B}) + P(B)\, P(\bar{A}) = \frac{9}{40} + \frac{10}{40} = \frac{19}{40}$

Hence percentage of cases in which they contradict each other is

$$= \frac{19}{40} \times 100 = 47.5\%$$

Illustration 21. In a certain town, male and female each form 50 per cent of the population. It is known that 20 per cent of the males and 5 per cent of the females are unemployed. A research student studying the employment situation selects an unemployed person at randos. What is the probability that the person so selected is (*a*) male (*b*) female?

Solution. The problem gives us the following probabilities as shown in the table below :

UNEMPLOYMENT DATA

	Unemployed	*Employed*	*Total*
Males	0.100	0.400	0.50
Females	0.025	0.475	0.50
Total	0.125	0.875	1.00

(a) P[Male/Unemployed] $= P(M/U) = \dfrac{P(M \cap U)}{P(U)} = \dfrac{0.10}{0.125} = 0.8$

(b) P[Female/ Unemployed] $= P(F/U) = \dfrac{P(F \cap U)}{P(U)} = \dfrac{0.025}{0.125} = 0.2$

Aliter

This illustration can also be solved by using Bayes' Theorem as shown below :

Given : $P(M) = 0.5,\ P(F) = 0.5,\ P(U/M) = 0.2,\ P(U/F) = 0.05$

(i) $$P(M/U) = \frac{P(U/M) \cdot P(M)}{P(U/M) \cdot P(M) + P(U/F) \cdot P(F)}$$

$$= \frac{0.2 \times 0.5}{0.2 \times 0.5 + 0.05 \times 0.5} = \frac{0.10}{0.125} = 0.8$$

Thus the probability that the unemployed person selected being a male is 0.8.

(ii) $$P(F/U) = \frac{P(U/F) \cdot P(F)}{P(U/M) \cdot P(M) + P(U/F) \cdot P(F)}$$

$$= \frac{0.05 \times 0.5}{0.2 \times 0.5 + 0.05 \times 0.5} = \frac{0.025}{0.125} = 0.2$$

Thus the probability that the unemployed person selected being a female is 0.2.

Illustration 22. A piece of equipment will function only when all the three components A, B and C are working. The probability of A failing during one year is 0.15, that of B failing is 0.05 and that of C failing is 0.10. What is the probability that the equipment will fail before the end of the year ?

Solution. The probability that component A does not fail during the year = 0.85

The probability that component B does not fail during the year = 0.95

The probability that component C does not fail during the year = 0.90.

Since the events are independent, therefore, the probability that all the three components do not fail = $0.85 \times 0.95 \times 0.90 = 0.727$.

Hence the probability that the equipment will fail before the end of the year

$$= 1 - 0.727 = 0.273$$

Illustration 23. Two computers A and B are to be marketed. A salesman who is assigned the job of finding customers for them has 60% and 40% chances respectively of succeeding in case of computer A and B. The computers can be sold independently. Given that he was able to sell at least one computer, what is the probability that computer A has been sold ?

(MBA, IGNOU, 2002;MBA, DU, 2002, 2006)

Solution. Let event A and B denote that the computer A and B are sold respectively,

Then, $P(A) = 0.60;\ P(B) = 0.40$

and $P(A \cap B) = P(A) \cdot P(B) = 0.60 \times 0.40 = 0.24$ [Independent events]

Probability of selling at least one computer is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.60 + 0.40 - 0.24 = 0.76$$

We are required to find out

$$P(A/A \cup B) = \frac{P(A \cup A \cap B)}{P(A \cup B)} = \frac{P(A)}{P(A \cup B)} = \frac{0.60}{0.76} = 0.7895.$$

Illustration 24. Explain whether or not each of the following claims could be correct :

(i) A businessman claims the probability that he will get contract A is 0.15 and that he will get contract B is 0.20. Furthermore, he claims that the probability of getting A or B is 0.50.

(ii) A market analyst claims that the probability of selling ten million kg. of plastic A or five million kg. of plastic B is 0.60. He also claims that the probability of selling ten million kg. of A and five million pounds of B is 0.45.

Solution. (i) Let event A and B denote the probability of getting contract A and B respectively.

Then, $P(A) = 0.15;\ P(B) = 0.20$ and $P(A \cup B) = 0.50$.

Probability of getting both the contracts is

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.15 + 0.20 - 0.50 = -0.15$$

Hence the claim of the businessman is wrong as the probability of getting both the contracts is negative

(*ii*) Let event A and B denote the selling of ten million of kg. of plastic A and five million kg. of plastic B respectively.

Then $P(A \cup B) = 0.60$ and $P(A \cap B) = 0.45$

Therefore $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

or $P(A) + P(B) = 0.60 + 0.45 = 1.05$

Since the sum of the probabilities is more than one, hence the claim of the market analyst is wrong.

Illustration 25. In a survey of 100 readers, it was found 40 read magazine A, 15 read magazine B, and 10 read both. What is the probability of a person reading at least one of the magazines ?

Solution. $P(A) = 0.40, \quad P(B) = 0.15 \quad P(A \cap B) = 0.10$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.40 + 0.15 - 0.10 = 0.45.$$

Therefore, the probability that a person read at least one magazine is 0.45.

Illustration 26. A bag contains 8 red and 5 white balls. The successive drawings of 3 balls are made such that (*i*) balls are replaced before the second trial. (*ii*) The balls are not replaced before the second trial. Find the probability that the first drawing will give 3 white and the second 3 red balls.

Solution. (*a*) *When the balls are replaced.*

Total number of balls in the bag = 13

Total number of red balls = 8

3 balls out of 13 can be drawn in ${}^{13}C_3$ ways.

3 white balls can be drawn out of 5 in ${}^{5}C_3$ ways.

3 red balls can be drawn out of 8 red in ${}^{8}C_3$ ways.

Since balls are replaced before the second draw, the total number of outcomes for both the draws remain the same, *i.e.*, ${}^{13}C_3$.

$$P(A) = \frac{{}^{5}C_3}{{}^{13}C_3} \text{ and } \quad P(B) = \frac{{}^{8}C_3}{{}^{13}C_3}$$

$$P(A \cap B) = P(A) \times P(B) = \frac{{}^{5}C_3}{{}^{13}C_3} \times \frac{{}^{8}C_3}{{}^{13}C_3} = \frac{140}{20449} = 0.0068.$$

(*b*) *When balls are not replaced.* When balls are not replaced, there would be no change in $P(A)$. Also the number of outcomes favourable to event B shall be ${}^{8}C_3$ but on the second draw the total number of outcomes would be ${}^{10}C_3$ (since 3 balls drawn are not replaced).

$$P(B/A) = P(B) = \frac{{}^{8}C_3}{{}^{10}C_3}$$

Hence $$P(A \cap B) = P(A) \times P(B) = \frac{{}^{5}C_3}{{}^{13}C_3} \times \frac{{}^{8}C_3}{{}^{10}C_3} = \frac{7}{429} = 0.0163.$$

Illustration 27. 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 fractions of defective output produced by the three plants are respectively at random 0.005, 0.008 and 0.010. If a pipe is selected from a day's total production and found to be defective, find out (*i*) from which plant for this defective pipe, the probability is highest. (*ii*) What is the probability that it came from the first plant? ***[MBA, IIT, Roorkee, 2004; M.B.A, Hyderabad Univ., 2005,M.B.A, Delhi Univ., 2006]***

Solution.

Let B_1 = Production volume of first plant.

B_2 = Production volume of second plant.

B_3 = Production volume of third plant.

A = a defective item.

$$P(B_1) = \frac{500}{3500} = \frac{1}{7} = 0.1428$$

$$P(B_2) = \frac{1000}{3500} = \frac{2}{7} = 0.2857$$

$$P(B_3) = \frac{2000}{3500} = \frac{4}{7} = 0.5714$$

$P(B_1)$, $P(B_2)$, $P(B_3)$ denotes the probability of selecting a unit from 1st, 2nd, 3rd plant respectively.

Now, calculating the joint probabilities

$$P(B_1 \cap A) = P(B_1) \times P(A/B_1) = 0.1428 \times 0.005 = 0.0007$$
$$P(B_2 \cap A) = P(B_2) \times P(A/B_2) = 0.2857 \times 0.008 = 0.00228$$
$$P(B_3 \cap A) = P(B_3) \times P(A/B_3) = 0.5714 \times 0.01 = 0.0057$$

Using Bayes' theorem, the required probabilities are :

$$P(B_1/A) = \frac{P(B_1 \cap A)}{P(B_1 \cap A) + P(B_2 \cap A) + P(B_3 \cap A)}$$
$$= \frac{0.0007}{0.0007 + 0.00228 + 0.0057} = \frac{0.0007}{0.00868} = 0.081$$

Similarly,
$$P(B_2/A) = \frac{0.00228}{0.00868} = 0.038$$
$$P(B_3/A) = \frac{0.0057}{0.00868} = 0.656$$

(*i*) As $P(B_3/A)$ has the highest probability we can say that it is most likely that the defective pipe has been drawn from the third plant.

(*ii*) The probability that it came from the first plant is given by $P(B_1/A)$ which is 0.081

Illustration 28. A market survey conducted in four cities pertained to preference for brand *A* soap. The responses are shown below :

	Delhi	*Kolkata*	*Chennai*	*Mumbai*
Yes	45	55	60	50
No	35	45	35	45
No opinion	5	5	5	5

(*i*) What is the probability that a consumer selected at random preferred brand *A* ?
(*ii*) What is the probability that a consumer preferred brand *A* and was from Chennai ?
(*iii*) What is the probability that a consumer preferred brand *A* given that he/she was from Chennai ?
(*iv*) Given that a consumer preferred brand *A*, what is the probability that he/she was from Mumbai ?

(MBA, Kumaon Univ., 1999; MBA, Delhi Univ., 2004, 2007)

Solution.

	Delhi	*Kolkata*	*Chennai*	*Mumbai*	*Total*
Yes	45	55	60	50	210
No	35	45	35	45	160
No opinion	5	5	5	5	20
Total	85	105	100	100	390

Let the event *A* denote that a consumer selected at random preferred brand *A*.

(*i*) $$P(A) = \frac{210}{390} = \frac{7}{13} = 0.5385$$

(*ii*) $$(A \cap C) = \frac{60}{390} = \frac{2}{13} = 0.1538$$

(*iii*) $$P(A/C) = \frac{P(A \cap M)}{P(M)} = \frac{60/390}{100/390} = \frac{3}{5} = 0.6$$

(*iv*) $$P(M/A) = \frac{P(B \cap A)}{P(A)} = \frac{50/390}{210/390} = \frac{5}{21} = 0.238$$

Illustration 29. In a locality, out of 5,000 people residing, 1,200 are above 30 years of age and 3,000 are females. Out of the 1,200 who are 30 years of age 200 are females. Suppose, after a person is chosen you are told that the person is female. What is the probability that she is above 30 years of age ?

(MBA, Delhi Univ., 2001)

Solution. Let event A denote that the person is above 30 years of age and event B that the person is a female.

Therefore $$P(B) = \frac{3000}{5000}, \quad P(A \cap B) = \frac{200}{5000}$$

Hence $$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{200/5000}{3000/5000} = 0.067$$

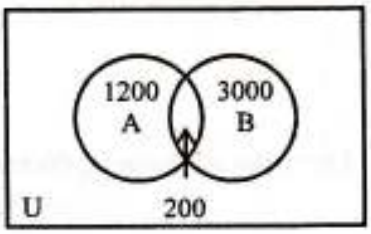


Illustration 30. Two factories manufacture the same machine part. Each part is classified as having either 0, 1, 2 or 3 manufacturing defects. The joint probability distribution for this is given below :

	Number of defects			
	0	1	2	3
Manufacturer A	0.1250	0.0625	0.1875	0.1250
Manufacturer B	0.0625	0.0625	0.1250	0.2500

(*i*) A part is observed to have no defects. What is the probability that it was produced by manufacturer A ?

(*ii*) A part is known to have been produced by manufacturer A. What is the probability that the part has no defects ?

(*iii*) A part is known to have two or more defects. What is the probability that it was manufactured by A ?

(*iv*) A part is known to have one or more defects. What is the probability that it was manufactured by B ?

Solution.

	Number of defects				
	0	1	2	3	
A	0.1250	0.0625	0.1875	0.1250	0.5000
B	0.0625	0.0625	0.1250	0.2500	0.5000
Total	0.1875	0.1250	0.3125	0.3750	1.0000

(*i*) $$P(A/\text{No defects}) = \frac{P(A \text{ and No defects})}{P(\text{No defects})} = \frac{0.1250}{0.1875} = 0.6667$$

(*ii*) $$P(\text{No defects}/A) = \frac{P(\text{ No defects and } A)}{P(A)} = \frac{0.1250}{0.5000} = 0.2500$$

(*iii*) $$P(A/\text{2 or more defects}) = \frac{P(A \text{ and 2 or more defects})}{P(\text{2 or more defects})} = \frac{0.3125}{0.6875} = 0.4545$$

(*iv*) $$P(B/\text{1 or more defects}) = \frac{P(B \text{ and 1 or more defects})}{P(\text{1 or more defects})} = \frac{0.4375}{0.8125} = 0.5385$$

Illustration 31. The probability that a new marketing approach will be successful is 0.6. The probability that the expenditure for developing the approach can be kept within the original budget is 0.50. The probability that both of these objectives will be achieved is 0.30.

What is the probability that at least one of these objectives will be achieved. For the two events described above, determine whether the events are independent or dependent. *(MBA, Delhi Univ., 2006)*

Solution : Let A denote the event that the new marketing approach will be successful and the event B denote the event that the expenditure for developing the approach can be kept within the original budget. Therefore, we are given

$$P(A) = 0.6, \quad P(B) = 0.5 \quad \text{also} \quad P(A \cap B) = 0.3$$

The probability that at least one of these objectives will be achieved is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.5 - 0.3 = 0.8$$

If events are independent

$$P(A \cap B) = P(A)\,P(B) = 0.6 \times 0.5 = 0.30$$

Which is same as given above. Hence events are independent.

Illustration 32. Of 1000 assembled components, 10 have a working defect and 20 have a structural defect. There is a good reason to assume that no component has both defects. What is the probability that randomly chosen component will have either type of defect ?

Solution. Let event A denote that the component has working defect and event B that the component has structural defect. Therefore, we are given

$$P(A) = \frac{10}{1000} = 0.01, \; P(B) = \frac{20}{1000} = 0.02$$

Also assuming that no component has both defects is given by $P(A \cap B)$. Therefore, the probability that the component will have either type of defect is given as :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.01 + 0.02 - 0.0 = 0.03.$$

Illustration 33. A study of Job satisfaction was conducted for four occupations. Cabin maker, lawyer, doctor, and systems analyst. Job satisfaction was measured on a scale of 0 – 100. The data obtained are summarized in the following cross tabulation.

Occupation	*Under 50*	*50 – 59*	*60 – 69*	*70 – 79*	*80 – 89*	*Total*
Cabin maker	0	2	4	3	1	10
Lawyer	6	2	1	1	0	10
Doctor	0	5	2	1	2	10
Systems Analyst	2	1	4	3	0	10
Total	8	10	11	8	3	40

(MBA, D.U., 2003)

(*i*) Develop a joint probability table.

(*ii*) What is the probability of one of the participants studied had a satisfaction score in the 80's ?

(*iii*) What is the probability of a satisfaction score in the 80's given the study participant was a doctor ?

(*iv*) What is the probability of one of the participants studied was a lawyer ?

(*v*) What is the probability of one of the participants was a lawyer and received a score under 50 ?

(*vi*) What is the Probability of a satisfaction score under 50 given a person is a lawyer ?

(*vii*) What is the probability of a satisfaction score of 70 or higher ?

(MBA, Delhi Univ., 2003)

Solution.

(*i*) Joint Probability table is :

Occupation	*Under 50*	*50 – 59*	*60 – 69*	*70 – 79*	*80 – 89*
Cabin maker	0.000	0.050	0.100	0.075	0.250
Lawyer	0.150	0.050	0.025	0.025	0.250
Doctor	0.000	0.125	0.050	0.025	0.250
Systems Analyst	0.050	0.025	0.100	0.075	0.250

(*ii*) P[Satisfaction score in the 80's] $= \frac{3}{40} = 0.075.$

(*iii*) P[Satisfaction score in the 80's/Doctor] $= \frac{P[SS\,(80)]}{P[D]} = \frac{2/40}{10/40} = \frac{1}{5} = 0.20.$

(*iv*) P [Lawyer] $= \frac{10}{40} = 0.25.$

(*v*) P [Lawyer and score under 50] $= \frac{P\,(\text{Lawyer} \cap \text{Score Under 50})}{P\,(\text{Score Under 50})} = \frac{6}{40} = 0.15.$

(*vi*) P [Score under 50/Lawyer] $= \frac{P\,[\text{Score under 50} \cap \text{Lawyer}]}{P\,[\text{Lawyer}]} = \frac{6/40}{10/40} = 0.6.$

(*vii*) P [Satisfaction score of 70 or higher]

$= P$ [score of 70 and above] + P [score of 80 and above]

$= \frac{8}{40} + \frac{3}{40} = \frac{11}{40} = 0.275.$

Illustration 34. Given the probabilities of three events, A, B and C are $P(A) = 0.35$, $P(B) = 0.45$ and $P(C) = 0.2$. Assuming that A, B and C have occurred, the conditional probabilities of another event, X, occurring are $P(X/A) = 0.8$, $P(X/B) = 0.65$ and $P(X/C) = 0.3$. Find $P(A/X)$, $P(B/X)$ and $P(C/X)$. *(MBA, IGNOU, 2007)*

Solution. We shall make use of Bayes' is theorem to solve this problem.

Given $P(A) = 0.35$, $P(B) = 0.45$, $P(C) = 0.20$

Also $P(X/A) = 0.80$, $P(X/B) = 0.65$, $P(X/C) = 0.30$

$$P(A/X) = \frac{P(A)\,P(X/A)}{P(A)\,P(X/A) + P(X/B)\,P(B) + P(X/C)\,P(C)}$$

$$= \frac{0.35 \times 0.8}{(0.35 \times 0.8) + (0.45 \times 0.65) + (0.2 \times 0.3)}$$

$$= \frac{0.28}{0.28 + 0.29 + 0.06} = \frac{0.28}{0.63} = 0.44$$

$$P(B/X) = \frac{P(B)\,P(X/B)}{P(A)\,P(X/A) + P(B)\,P(X/B) + P(C)\,P(X/C)}$$

$$= \frac{(0.45)\,(0.65)}{(0.35 \times 0.8) + (0.45)\,(0.65) + (0.2)\,(0.3)}$$

$$= \frac{0.29}{0.28 + 0.29 + 0.06} = \frac{0.29}{0.63} = 0.46$$

$$P(C/X) = \frac{P(C)\,P(X/C)}{P(A)\,P(X/A) + P(B)\,P(X/B) + P(C)\,P(X/C)}$$

$$= \frac{(0.2)\,(0.3)}{(0.35 \times 0.8) + (0.45 \times 0.65) + (0.2 \times 0.3)}$$

$$= \frac{0.06}{0.28 + 0.29 + 0.06} = \frac{0.06}{0.63} = 0.09.$$

PROBLEMS

1-A : Answer the following questions, each question carries **one** mark:

(*i*) State the addition law of probability.

(*ii*) What do you mean by the term conditional probability ? *(MBA, Hyderabad Univ., 2006)*

(*iii*) A bag contains 7 red balls and 5 white balls. 2 balls are drawn at random. What is the probability that all of them are red ?

(*iv*) Define the term probability.

(*v*) What are independent events ?

(*vi*) Explain with examples the rule of addition in theory of probability. *(MBA, Madurai-Kamaraj Univ., 2003)*

(*vii*) What is meant by mutually exclusive events ?

(*viii*) Explain Bayes' theorem with the help of a suitable example. *(MBA, Hyderabad Univ., 2005)*

1-B : Answer the following questions, each question carries **four** marks:

(*i*) Three balls are drawn at random from a basket containing 6 blue and 4 red balls. What is the chance that two balls are blue and one is red ?

(*ii*) What is the probability that a leap year selected at random will contain 53 Sundays ? *(M.Com., M.K. Univ., 2008)*

(*iii*) What do you mean by probability ? Explain the importance of probability. *(M.A. Econ., Madras Univ., 2008)*

(*iv*) What are the basic laws of probability ? *(MBA, Madras Univ., 2003)*

(*v*) State and prove the addition law of probability.

2. Explain what do you understand by the term probability. Discuss its importance in managerial decision-making. *(MBA, Delhi Univ., 2007)*

3. Describe briefly the various schools of thought on probability. How does the concept of probability help decision-maker to improve his decisions ?

4. (a) Explain the various approaches to probability. Are they contradictory ?
   (b) Examine critically the different schools of thought on probability.

(*MBA, Kumaon Univ., 2000*)

5. Explain with examples the rules of Addition and Multiplication in theory of probability.
6. Give the classical and statistical difinitions of probability and state the relationship, if any, between the two definitions.
7. State and prove the addition and multiplication theorems of probability.
8. (a) Explain with the help of an example the concept of conditional probability.
   (b) Explain the concept of conditional probability and Bayes' theorem. (*MBA, IGNOU, 2002*)
9. Explain the difference between :
   (i) Simple probability and conditional probability.
   (ii) Independent event and mutually exclusive event.
10. Define independent and mutually exclusive events. Can two events be mutually exclusive and independent simultaneously? Support your answer with an example. (*MBA, Delhi Univ., 2007*)
11. When are two events said to be independent in the probability sense ? Give examples of dependent and independent events.
12. (a) Explain the concept of probability following the experimental frequency approach.
    (b) What do you understand by conditional probability ? If Prob. $(A + B)$ = Prob. (A) + Prob. (B), are the two events A and B statistically independent ?
13. Write an essay on prior and posterior probabilities and Bayes' theorem and also show how Bayes' theorem can be extended in the case of n events.
14. (a) Make up a realistic problem from your area of interest to illustrate the use of Bayes' theorem.
    (b) State the multiplicative theorem of probability. How is the result modified when the events are independent ? (*M.Com., Delhi Univ., 2009*)
15. The personnel manager of a large manufacturing firm finds that 15 per cent of the firm's employees are junior executives and 25 per cent of the firm's employees are MBAs. He also discovers that 5 per cent of the firm's employees are both junior executives and MBAs. What is the probability of selecting a junior executive if the selection is confined to MBAs ?
    [0.20]
16. A company learned that inventory shortages were associated with a loss of goodwill with a probability 0.10. The company also knew that a loss of goodwill from all causes occurred with a probability of 0.15. What is the probability of an inventory shortage, given a loss of goodwill ?
    [0.67]
17. An article manufactured by a company consists of two parts A and B. In the process of manufacture of part A, 9 out of 100 are likely to be defective. Similarly, 5 out of 100 are likely to be defective in the manufacture of part B. Calculate the probability that the assembled part will not be defective.
    [0.8645]
18. A candidate is selected for interview for three posts. For the first post there are 3 candidates, for the second there are four and for the third there are two candidates. What are his chances of getting at least one post ?
    [0.75]
19. An investment firm purchases 3 stocks for one week trading purposes. It assesses the probability that the stocks will increase in value over the week as 0.8, 0.7 and 0.6 respectively. What is the chance (*i*) all three stocks will increase and (*ii*) at least 2 stocks will increase ? (Assume that the movements of these stocks are independent.)
    [(*i*) 0.336, (*ii*) 0.788.]
20. A company has two plants to manufacture scooters. Plant 1 manufactures 80% of the scooters and plant 2 manufactures 20%. At Plant 1, 85% scooters are rated as standard quality. At plant 2, only 65% scooters are rated as standard quality.
    (*i*) What is the probability, that a customer obtains a standard quality scooter if he buys a scooter from the company ?
    (*ii*) What is the probability, that the scooter came from plant 1, if it is known that the scooter is of standard quality ?
    [(*i*) 0.81, (*ii*) 0.84]
21. 10% of the employees of a certain company have been to public school. Of these, 30% hold administrative positions. Of these that have not been to public school, 30% hold administrative positions. If an employee is selected at random from the administrative staff, what is the probability that he was educated in a public school ? (*MBA, HPU, 2009*)
22. A factory produces a mechanism which consists of three independently manufactured parts. It is known that 1 per cent of part one, 4 per cent of part two and 3 per cent of part three are defective. What is the probability that a complete mechanism is not defective ?
    [0.9218]

23. A manager has two assistants and he bases his decision on information supplied independently by each of them. The probability that he makes a mistake in his thinking is 0.005. The probability that an assistant gives wrong information is 0.3. Assuming that the mistakes made by the manager are independent of the information given by the assistants, find the probability that he reaches a wrong decision. *(MBA, DU, 2001)*

[0.5122]

24. A piece of electronic equipment has two essential parts, A and B. In the past, part A has failed 40% of the time; part B 50% of the time. Parts *A* and *B* operate independently. Assume that both parts must operate to enable the equipment to function. What is the probability that the equipment will function ?

[0.30]

25. Three groups of workers contain 3 men and 1 woman, 2 men and 2 women, and 1 man and 3 women, respectively. One worker is selected at random from each group. What is the probability that the group selected consists of 1 man and 2 women ?

[0.4063]

26. Two sets of candidates are competing for the position on the Board of Directors of a company. The probability is that the first and second sets will win 0.6 and 0.4 respectively. If the first set wins, the probability of introducing a new product is 0.8, and the corresponding probability if the second set wins is 0.3. What is the probability that the new product will be introduced?

[0.60]

27. There are three cars, *A, B* and *C*. Car *A,* contains two males, car *B* contains one male and one female, and car *C* contains two females. If one of these cars is selected at random, and one person is observed to be male, what is the probability that the other person in that car is male ?

[2/3]

28. A salesman has a 60 per cent chance of making a sale to each customer. The behaviour of successive customers is independent. If two customers *A* and *B* enter, what is the probability that the salesman will make a sale to *A* or *B*?

[0.84] *(MBA, DU, 1998)*

29. A factory produces certain types of output by three machines. The respective daily production figures are : Machine $A = 3{,}000$ units; Machine $B = 2{,}500$ units; and Machine $C = 4{,}500$ units. Past experience shows that 1 per cent of the output produced by machine *A* is defective. The corresponding fractions of defectives for the other two machines are 1.2 and 2 per cent respectively. An item is drawn at random from the day's production run and is found to be defective. What is the probability that it comes from the output of (*i*) Machine *A;* (*ii*) Machine *B*; and (*iii*) Machine *C* ?

[(*i*) 0.2; (*ii*) 0.2; (*iii*) 0.6]

30. In a bolt factory machines *A, B* and *C* manufacture respectively 25%, 35% and 40% of the total of their output 5,4 and 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* ? *(MBA, Delhi Univ., 2010)*

[$A = 0.37$; $B = 0.40$; $C = 0.23$]

31. In a factory manufacturing pens. Machines *X, Y* and *Z* manufacture 30, 30 and 40 per cent of the total production of pens respectively.

Of their output 4, 5 and 10 per cent of the pens are defective. If one pen is selected at random it is found to be defective what is the probability that it is manufactured by machine Z ?

[0.6639] *(MBA, UP Tech. Univ., 200*

32. The probability that India wins a cricket test match against Pakistan is, given to be 1/3. If India and Pakistan play six te matches, what is the probability that :

(*i*) India will lose all the six test matches ?

(*ii*) India will win at least one test match ? *(MBA, Delhi Univ., 200*

[(*i*) 0.088; (*ii*) 0.912]

33. Three persons *A, B* and *C* are being considered for the appointment as Vice-Chancellor of a university whose chances of be selected for the post are in the proportion 4:2:3 respectively. The probability that *A*, if selected, will introduce democratisat in the University Structure is 0.3 and the corresponding probablities for *B* and *C* doing the same are respectively 0.5 and What is the probability that democratisation would be introduced in the University ? *(MBA, DU, 2(*

[0.511]

34. 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
(*a*) both of them will be selected,
(*b*) only one of them will be selected, and
(*c*) none of them will be selected ?
[(*a*) 1/35, (*b*) 10/35, (*c*) 24/35] *(M.Com., Madurai Kamaraj Univ., 2003)*

35. A hotel gets cars for its guests from three rental agencies, 20 per cent from agency *X*, 40 per cent from agency *Y* and 40 per cent from *Z*. If 14 per cent of the cars from *X*, 4 per cent from *Y* and 8 per cent from *Z*, need tune-ups, what is the probability that car needing a tune-up is delivered to one of the hotel's guests ? *(MBA, DU, 1999)*
[7.6%]

36. The odds against student *X* solving a Business statistics problem are 8 : 6 and odds in favour of student *Y* solving the same problem are 14:16. (*i*) What is the chance that the problem will be solved if both try ? (*ii*) What is the probability that they both, working independently of each other, solve the problem ? (*iii*) What is the probability that neither solves the problem ?
[(*i*) 0.6952; (*ii*) 0.2; (*iii*) 0.3048] *(MBA, Delhi Univ., 2004, 2007)*

37. In a certain government office there are 400 employees; there are 150 men, 276 University graduates, 212 married persons, 94 male university graduates, 151 married university graduates, 119 married men, 72 married male university graduates. Find the number of single women who are not university graduates ?
[54]

38. A lot of vacuum tubes contains 1000 tubes. 10 of which have a defective grid and no other defects and 20 of which have both a defective grid and defective heating element. A tube is drawn at random from the lot and we are told that it has defective grid. What is the probability that it also has a defective heating element ? What model did you use in computing this probability ?

39. Probability that a man will be alive 25 years hence is 0.3 and the probability that his wife will be alive 25 years hence is 0.4. Find the probability that 25 years hence (*i*) both will be alive (*ii*) only the man will be alive (*iii*) only the woman be alive and (*iv*) at least one of them will be alive.

40. A convention begins with an evening lecture, attended by 60% of the delegates. The following morning lecture is attended by 10% of the delegates. Seventy per cent of those attending this session had attended the previous evening session.
(*i*) What is the probability that a randomly selected delegate attended both the sessions ?
(*ii*) What is the probability that a delegate who attended the evening session also attended the following morning session ?
(*iii*) What is the probability that a delegate selected at random attended at least one of the two sessions ?
(*iv*) Are attendances at the two sessions statistically independent ?

41. Mr. Ram speaks the truth in 3 out of 4 times, while Mr. Shyam speaks the truth in 4 out of 5 times. Find the probability that they will contradict each other in stating the fact.
[0.35]

42. Two union leaders and 10 directors of a company sit randomly around a round table to decide upon the wage hike as demanded by the union. Find the probability that there will be exactly three directors between the two union leaders.

43. Assume we have three boxes which contain red and black balls as follows :

Box 1	—	3 red and 7 black
Box 2	—	6 red and 4 black
Box 3	—	8 red and 2 black

A ball is drawn from box 1; if it is red, 2nd ball is drawn from box 2. If the 1st ball drawn from box 1 is black, 2nd ball is drawn from box 3.
(*i*) What is the probability that the two balls are red ?
(*ii*) What is the probability that one ball is red and another ball is black ?

44. Explain whether or not each of the following claims could be correct :
(*a*) A supplier claims that the long-run fraction of the resistors he produces which are defective is 0.001. In one lot of 10,000 resistors obtained from the supplier 30 defectives were discovered.
(*b*) A plant engineer claims the probability that machine will not fail in a one month period is 0.20, the probability that it will fail exactly once is 0.50, the probability that it will fail twice is 0.30 and the probability that it will fail more than twice is 0.30.
(*c*) A market analyst claims that the probability that sales of less than 4 million pounds in the next year is 0.3, of sales between 4 and 6 million pounds is 0.4 and sales of more than 6 million pounds is 0.2.

45. A production process which turns out transistors has a long-run fraction defective of 0.005. A testing device is used to check each transistor produced. It has been found that the device always indicates that a defective is indeed defective, but for about 1 in every 100 transistors produced it indicates that a good transistor is defective. If the device indicates that a given transistor is defective, what is the probability that it is actually defective ?

46. A certain production process produces items that are 10 per cent defective. Each item is inspected before being supplied to customers but the inspector incorrectly classifies an item 10 per cent of the time. Only items classified as good are supplied. If 820 items in all have been supplied, how many of them are expected to be defective ?
[10]

47. A market research firm is interested in surveying certain attitudes in a small community. There are 1,250 households broken down according to income, ownership of a telephone and ownership of a TV.

	Households with annual income of Rs. 3,00,000 or less		*Household with annual income above Rs. 3,00,000*	
	Telephone Subscriber	*No Telephone*	*Telephone Subscriber*	*No Telephone*
Own TV set	270	200	180	100
No TV set	180	100	120	100

(*i*) What is the probability of obtaining a TV owner in drawing at random ?
(*ii*) If a household has annual income over Rs. 3,00,000 and is a telephone subscriber, what is the probability that he has a TV?
(*iii*) What is the conditional probability of drawing a household that owns a TV given that the household is a telephone subscriber.
(*iv*) Are the events 'ownership of a TV' and 'telephone subscriber' statistically independent ? Comment.
[(*i*) 0.6, (*ii*) 0.6, (*iii*) 0.6, (*iv*) yes]

48. Past surveys show that 40% of the officers at a certain industry own cars. Suppose six officers are selected at random from this industry (with replacement).
(*a*) What is the probability that exactly four will own cars ?
(*b*) What is the probability that at least one will own a car ?
(*c*) What is the theoretical mean of the probability distribution under consideration ?

49. A survey reports that 80% of the population is married and 55% is male. What is the least possible percentage of married men and of married women ?

50. Consider a family with two children. Assuming that each child is as likely to be a boy as it is to be a girl, what is the conditional probability that both children are boys given that (*a*) the elder child is a boy, (*a*) at least one of the children is a boy?

51. A man goes for fishing for the first time. He has three types of bait, only one of which is correct for the type of fish he intends to try. The probability that he will catch a fish if he uses correct bait is 1/3. If he uses the wrong bait, his chances of catching a fish are 1/5.
(*a*) What is the probability that he will catch a fish ?
(*b*) Given the man caught a fish, what is the probability that he used correct type of bait ?

52. If a machine is correctly set up, it will produce 90% acceptable items. If it is incorrectly set up, it will produce 40% acceptable items. Past experience shows that 80% of the set-ups are correctly done. If after a certain set-up the machine produces 2 acceptable items as the first 2 pieces, find the probability that the machine is correctly set up.

53. Consider two events A and B such that $P(A) = 1/8$, $P(A/B) = 1/4$ and $P(B/A) = 1/6$. Examine the following statements and comment on the validity of each of these :
(*i*) A and B are independent.
(*ii*) A and B are mutually exclusive.
(*iii*) Occurrence of A implies that of B.
(*iv*) $P(A/B) = 0.5$.

54. If a pair of dice is thrown, find the probability that the sum is neither 7 nor 11.
[7/9]

55. An investment consultant predicts that the odds against the price of a certain stock will go up during the next week are 2 : 1 and the odds in favour of the price remaining the same are 1 : 3. What is the probability that the price of the stock will go down during the next week ?

56. (*a*) What is the probability that a leap year selected at random will contain either 53 Thursdays or 53 Fridays ?
[3/71]
(*b*) What is the probability that a leap year selected at random will contain 53 sundays?
(M. Com., Madurai-Kamaraj Univ., 2002)

57. A product is assembled from three components X, Y and Z and the probability of these components being defective is 0.01, 0.02 and 0.05. What is the probability that the assembled product will not be defective ? *(MBA, DU, 2002)*
[0.922]

58. According to a survey, the probability that a family owns 2 cars if their annual income is greater than Rs. 15,000 is 7. Of the households surveyed, 50 per cent had income over Rs. 15.000 and 40 per cent had 2 cars. What is the probability that a family has 2 cars and an income over Rs. 15,000 a year ?

59. A box contains 8 red, 3 blue and 9 green balls. If three balls are drawn at random, determine the probability that :
(*i*) all 3 are red ; (*ii*) all 3 are blue ; (*iii*) at least 1 is blue ; (*iv*) 2 are red and 1 green
(*v*) 1 of each colour ; and (*vi*) the balls are drawn in the order red, blue and green colours.

60. A problem in statistics is given to the three students X, Y and Z, whose chances of solving it are 1/3, 1/4, 2/5 respectively. What is the probability that the problem will be solved ?
[7/10]

61. A survey of readership of a certain investment magazine indicates that the proportion of male readers over 40 years is 0.02. The proportion of male readers under 40 is 0.07. What is the probability of a reader being a male ?

62. The cricket team of a University played four matches in Inter-University cricket matches. The captain of the team observed the practice of calling out "Heads" every time when the toss was made. What is the probability of his winning the toss in all the four matches ?
How would the probability be affected if the Captain had made a practice of tossing coin privately before calling out "Head" or "Tail" on each occasion ?

63. A sample of 3 items is selected at random from a box containing 12 items of which 3 are defective. Find the possible number of defective combinations of the said 2 selected items along with probability of a defective combination.

64. In an examination, 30% of the students have failed in Mathematics, 20% of students have failed in Chemistry and 19% have failed in both Mathematics and Chemistry. A student is selected at random.
(*i*) What is the probability that the student has failed in Mathematics if it is known that he has failed in Chemistry ?
(*ii*) What is the probability that the student has failed in Mathematics or in Chemisty ?

65. A company has four production sections, *viz.*, S_1, S_2, S_3 and S_4 which contribute 30%, 20%, 22% and 28% respectively to the total output. It was observed that these sections respectively produced 1%, 2%, 3% and 4% defective units. If a unit is selected at random and found to be defective, what is the probability that the unit so selected has come from either section one or section four ? *(MBA, GGSIPU, 2000; MBA, Delhi Univ., 2004)*

66. A factory has two machines. The empirical evidence has established that Machines (*i*) and (*ii*) produce 30% and 70% of the output respectively. It has also been established that 5% and 1% of the output produced by these machines respectively was defective. A defective item is drawn at random. What is the probability that the defective item was produced by machine (*i*) and (*ii*) ?
[(*i*) 0.682 (*ii*) 0.318]

67. It is believed that in 100 cases of income tax raids, and undisclosed income of more than Rs. 1 lakh is selected. What is the probability that the income tax office will have to make at most 10 raids until the first case of undisclosed income of more than Rs. 1 lakh is detected.

68. A box contains 10 white, 7 black and 3 green balls. 2 balls are drawn at random. Find out the probability that :
(*i*) both are white.
(*ii*) one is white and another is green.
(*iii*) one is black and another is green.
Find the probabilities in case of without replacement.

69. Project VIJAY, NCSO, INDIA sums its operations on 10 computers which may need repairs from time to time during the day. Three of these computers are old, each having a probability of 1/11 of needing repair during the day and seven are new, having corresponding probability of 1/ 21.
Assuming that no computer needs repair twice on the same day, determine the probabilities that on a particular day.
(*i*) just 2 old and no new computers need repair.
(*ii*) if just 2 computers need repair, they are of same type. *(MBA, IGNOU, 2000)*

70. A consignment of 20 picture tubes contain 5 defectives. Two tubes are selected one after another at random. Find the probability that both are defective assuming (*a*) the first is replaced before drawing the second, and (*b*) the first is not replaced.

71. A manager has drafted a scheme for the benefit of employees. To get an idea of the support for the scheme, he randon polls literate workers (L) and illiterate workers (I). He polls 30 of each group with the following results :

Opinion For Scheme	*L*	*I*
Strongly Support	9	10
Mildly Support	11	3
Undecided	2	2
Mildly oppose	4	8
Strongly oppose	4	7

(*a*) What is the probability that a literate worker selected randomly from the polled group mildly supports the scheme
(*b*) What is the probability that a worker (literate or illiterate) selected randomly from the polled group strongly or mild supports the scheme ?
(MBA, IGNOU, 200

72. Three institutions (*A, B,* and *C*) train students for MBA entrance test. They train in the proportion 25 per cent (*A*), 35 p cent (*B*) and 40 per cent (*C*) of the trained candidates, for *A, B, C*; 5 per cent, 4 per cent and 2 per cent are successful in t entrance test respectively.
A candidate is selected at random and found to be successful in the entrance, find the probability, that he was trained by *B,* or *C*. What is the probability of average success in the MBA entrance ? *(MBA, Bharathidasan Univ., 200*

73. A man either drives a car or catches a train to go to office each day. He never goes 2 days in a row by train but if he drive one day, then the next day he is just as likely to drive again as he is to travel by train. Now suppose that on the first day the week, the man tossed a fair dice and drove to work if and only if a 6 appeared. Find the probability that he takes a tra on the third day and also the probability that he drives to work in the long run.

74. A machine goes out of order, whenever a component fails. The failure of this part follows a Poisson process with a mean ra of 1 per week. Find the probability that 2 weeks have elapsed since last failure. If there are 5 spare parts of this compone in an inventory and that the next supply is not due in 10 weeks, find the probability that the machine will not be out of ord in the next 10 weeks. *(B.E./ B. Tech., Madras Univ., 200*

75. A manufacturing firm produces pipes in two plants I and II with daily production 1,500 and 2,000 pipes respectively. Th fraction of defective pipes produced by the two plants I and II are 0.006 and 0.008 respectively. If a pipe selected at rando from the day's production is found to be defective, what is the probability that it has come from plant I, plant II ?
(MBA, Bharathidasan Univ., 200

76. In a survey of MBA students, the following data were obtained on students' first reason for application to one business school

	Reason for application		
	School quality	*Placement*	*Other*
Enrollment	421	393	76
Status	400	593	46

(*i*) If a student goes full time, what is the probability that the school quality is the first reason for choosing a business school
(*ii*) If a student goes part time, what is the probability that the school quality is the first reason for choosing a busines school ?
(*iii*) Let *A* be the event that a student is full time and let *B* be the event that the student lists the school quality as the first reason for applying. Are events *A* and *B* independent? Justify your answer. *(MBA, Delhi Univ., 2006)*

77. Logic Dynamic Ltd., a computer manufacturing firm, receives shipment of parts from two different suppliers. Supplier *A* supplies the 70% of the total parts and the remaining 30% is supplied by supplier *B*. The historical quality levels of these two suppliers are shown in the following table:

	Good parts (%)	*Defective parts (%)*
Supplier *A*	95	5
Supplier *B*	90	10

(*i*) A part is randomly selected from the firm's inventory, and it is found to be defective, what is the probability that it is supplied by the supplier *A*?
(*ii*) A part is randomly selected from the firm's inventory, and it is found to be good, what is the probability that it is supplied by the supplier *B* ?
(MBA, Delhi Univ., 2009)

78. A vending machine at MacDonald's fast food restaurant automatically pours soft drinks into cups. The amount of soft drinks dispensed into a cup is normally distributed with mean 9.6 oz and standard deviation 0.6 oz.
(*i*) Estimate the probability that the machine overflow an 10 oz cup.
(*ii*) Estimate the probability that the machine will not overflow an 10 oz cup.
(*iii*) The machine has just been loaded with 745 cups. How many of these do you expect will overflow when served ?
(MBA, Delhi Univ., 2009)

79. A manufacturing company has plants in India produces 40% of the total output with 10% defective rate and the average cost of producing $12.50 per unit. If a single unit is found to be defective, what is the probability it has come from India ?

*****

12

Probability Distributions

INTRODUCTION

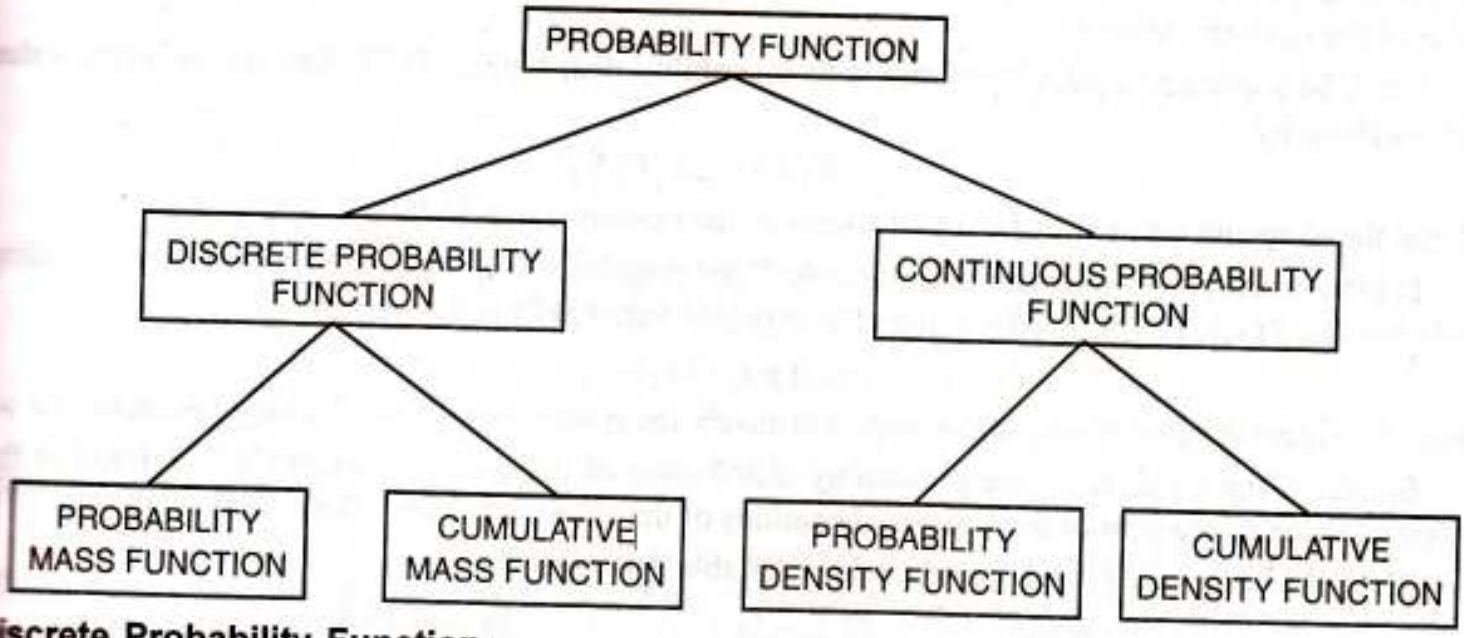
In this chapter, the basic concepts of probability distributions are discussed. In a probability distribution, the variables are distributed according to some definite probability function. We shall discuss some of the most important probability distributions in this chapter. These distributions from their historical interest as well as intrinsic importance, occupy a place of great prominence in business decision-making. Some of the concepts that will facilitate understanding of the topic are given below:

Random Variable

A random variable is a variable which takes specified values with specified probabilities. The probabilities are specified by the way in which the random experiment is conducted and the way in which the random variable is defined and observed on the random experiment. We shall use capital letters to denote a random variable and the corresponding small letters to represent any specific value of the random variable.

Probability Function

If the function permits us to compute the probability for any event that is defined in terms of value of the random variable, then the function is called a probability function. Just as there are discrete and continuous random variables, so there are discrete and continuous probability functions. To emphasize this distinction, we shall draw the diagram as follows:



Discrete Probability Function

A probability function for a discrete random variable is called a discrete probability function since the domain of the function is discrete.

Probability Mass Function

A probability function that specifies the probability that any single value of discrete random variab will occur is called a probability mass function (abbreviated as p.m.f.). If $f(x)$ is the probability ma function of the random variable X, then $f(x) = P(X = x)$ has the following properties :

(*i*) $f(x) \geq 0$ for all values of X; and

(*ii*) $\Sigma f(x) = 1$

Cumulative Mass Function

If X is a discrete random variable with p.m.f. $f(x)$, its cumulative mass function (abbreviated c.m.f.) specifies the probability that an observed value of X will be no greater than x. That is, if $F(x)$ a c.m.f. and $f(x)$ is a p.m.f., then $F(x) = P(X \leq x) = f(X \leq x)$.

Continuous Probability Function

A probability function for a continuous random variable is called a continuous probability functi since the domain of the function is continuous.

Probability Density Function

For a continuous random variable, the corresponding function $f(x)$ is called a probability dens function (abbreviated as p.d.f.). Unlike a p.m.f., a p.d.f. does not specify probabilities for specific individu values of the random variable.

Cumulative Density Function

Corresponding to the cumulative mass function of a discrete random variable, the cumulative dens function (abbreviated as c.d.f.) of a continuous random variable specifies the probability that an observ value of X will be no greater than x.

Expected Value and Variance

The probability distribution provides a model for the theoretical frequency distribution of a rand variable and hence must possess a mean, variance and other descriptive measures associated with t theoretical population which it represents. The average value of a random variable is called the expec value of the random variable.

Let X be a discrete random variable with probability distribution, $P(X)$, then the expected va $E(X)$ is given by

$$E(X) = \Sigma X . P(X)$$

where, the elements are summed over all values of the random variable X.

In other words, if a discrete random variable X has possible values $x_1, x_2, ..., x_n$, with correspond probabilities $P(x_1), P(x_2), ..., P(x_n)$ then the expected value $E(X)$ is defined as

$$E(X) = x_1 P(x_1) + x_2 P(x_2) + ... + x_n P(X_n) = \mu$$

Thus, the expected value of random variable X is merely the arithmetic mean which may be denoted by

Similarly, the variance of the probability distribution of the random variable X is defined as expected value of the sum of the squared deviations of the values of X from their mean.

Thus, the variance of the discrete random variable X is given by

$$\text{Var}(X) = \sigma^2 = E[X - E(X)]^2 = \Sigma[X - E(X)]^2 P(X)$$
$$= E(X^2) - [E(X)]^2.$$

The standard deviation, σ, is the square root of the variance.

Properties of Expected Value and Variance

There are several important properties of expected value and variance which allow computational shortcuts :

1. The expected value of a constant c is equal to the constant.

$$E(c) = c$$

2. The expected value of the product of a constant c and a random variable X is equal to the constant times the expected value of the random variable.

$$E(cX) = cE(X)$$

3. The expected value of the sum of a random variable X and a constant c is the sum of the expected value of the random variable and the constant.

$$E(X + C) = E(X) + c$$

4. The expected value of the product of two independent random variables is equal to the product of their individual expected values.

$$E(XY) = E(X)\,E(Y)$$

5. The expected value of the sum of the two independent random variables is equal to the sum of their individual expected values.

$$E(X + Y) = E(X) + E(Y)$$

6. The variance of the product of a constant and a random variable X is equal to the constant squared times the variance of the random variable X.

$$\text{Var}(cX) = c^2\,\text{Var}(X)$$

7. The variance of the sum of two independent random variables equal the sum of their individual variances. Also, the variance of the difference of two independent random variables equal the sum of their individual variances.

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) = \text{Var}(X - Y)$$

Illustration 1. Anil company estimates the net profit on a new product, it is launching, to be Rs. 3,000,000 during the first year if it is 'successful' Rs, 1,000,000 if it is moderately successful and a loss of Rs. 1,000,000 if it is 'unsuccessful'. The company assigns the following probabilities to first year prospects for the product, successful : 0.15, moderately successful: 0.25, and unsuccessful: 0.60. What are the expected value and standard deviation of first year net profit for the product? *(MBA, DU, 2003)*

Solution. The probability distribution of net profit (X) of the new product in the first year is given to be

Profit (in million Rs.) X	3	1	–1
Probability $P(X)$	0.15	0.25	0.60

Therefore, expected value of profit is given by

$$E(X) = \Sigma X P(X)$$
$$= 3 \times 0.15 + 1 \times 0.25 + 1 \times 0.60$$
$$= 0.10 \text{ million Rs.} = \text{Rs. } 1{,}00{,}000.$$
$$E(X^2) = \Sigma X^2 P(X)$$
$$= 9 \times 0.15 + 1 \times 0.25 + 1 \times 0.6 = 2.20 \text{ million Rs.}$$
$$\text{Var}(X) = E(X^2) - [E(X)]^2 = 2.20 - (0.10)^2 = 2.19$$
$$\text{s.d.} = \sigma = \sqrt{2.19} = 1.48 \text{ (Million Rs.)}$$

Binomial Distribution

The Binomial distribution is also known as the outcome of Bernoulli process and is associated with the name of Jacob Bernoulli. A Bernoulli process is a random process in which :

(a) the process is performed under the same conditions for a fixed and finite number of trials, say, n.

(b) each trial is independent of other trials, *i.e.*, the probability of an outcome for any particular trial is not influenced by the outcomes of the other trials.

(*c*) each trial has two mutually exclusive possible outcomes, such as "success" or "failure", "good" or "defective", "yes" or "no", "hit" or "miss", and so on. The outcomes are usually called success and failure for convenience.

(*d*) the probability of success, p, remains constant from trial to trial (so is the probability of failure q, where, $q = 1 - p$).

These conditions are satisfied if we toss a coin, say, five times. Suppose we are interested in finding the probability of obtaining exactly two heads.

Let us designate head as a success and tail as a failure with corresponding probabilities p and respectively. Find the probability of getting exaclty two heads of five tosses of a fair coin.

Suppose that one of the sequence of outcomes of five tosses of a fair coin showing two heads is :

HTHTT

The probability of this specific sequence of outcome is found by means of a multiplication rule of probability and is given by

$$pqpqq = p^2q^3$$

Although the resulting probability of obtaining the specific sequence of outcomes in the order shown, we are not interested in the order of occurrence of the successes and failures. Rather, we are interested in the probability of the occurrence of exactly two successes out of five tosses of a coin. In addition to the sequence shown above (call it sequence number 1), two successes and three failures could also occur in any one of the additional sequences shown as follows. Each of the sequences has the same probability of occurring, p^2q^3.

Sequence Number	*Sequence*	*Probability*
2	HHTTT	p^2q^3
3	HTTTH	p^2q^3
4	TTTHH	p^2q^3
5	TTHHT	p^2q^3
:	:	:
:	:	:
10	:	p^2q^3

A single sample of five tosses will yield only one sequence of successes and failures. The question then to be answered is: What is the probability of getting sequence number 1 or sequence number 2... or sequence number 10? For finding the answer, the addition rule of probability is used to calculate the sum of the individual probabilities.

To get this, we multiply p^2q^3 by 10, *i.e.*, $10\,p^2q^3$.

Here, $p = 0.5$ and $q = 0.5$

Therefore, the answer is

$$10\,(0.5)^2\,(0.5)^3 = 10 \times 0.25 \times 0.125 = 0.3125$$

As the size of the tosses increases, it becomes more and more difficult to list the number of sequences. An easy method of counting them is required.

We know that the number of combinations of n things taken x at a time is given by

$${}^nC_x = \frac{n!}{x!(n-x)!}$$

In our example, $n = 5, x = 2.$

Then $${}^5C_2 = \frac{5!}{2!3!} = 10.$$

Therefore, the general model for specifying the probability of obtaining exactly x successes in a given number of n, Bernoulli trials is given by

$$f(x) = P[X = x] = {}^nC_x p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \ldots\ldots, n.$$

where, p = the probability of a success on a single Bernoulli trial
n = the number of Bernoulli trials
x = the number of successes in n trials.

This formula for the probability distribution of the number of successes in series of Bernoulli trials is called the Binomial probability distribution. It gives the probability of obtaining exactly x successes and $(n - x)$ failures in n Bernoulli trials. The Binomial distribution has been extensively tabulated for different values of x and n (See Appendix).

Number of successes x	*Probability f(x)*
0	${}^nC_0 p^0 q^{n-0} = q^n$
1	${}^nC_1 p^1 q^{n-1} = nq^{n-1}p$
2	${}^nC_2 p^2 q^{n-2} = \frac{n(n-1)}{2} q^{n-2} p^2$
:	:
:	:
x	${}^nC_x p^x q^{n-x}$
:	:
:	:
n	${}^nC_n p^n q^{n-n} = p^n$

The binomial distribution satisfies the two essential properties of probability distribution, *viz.*, (*i*) $f(x) \geq 0$; and (*ii*) $\Sigma f(x) = 1$.

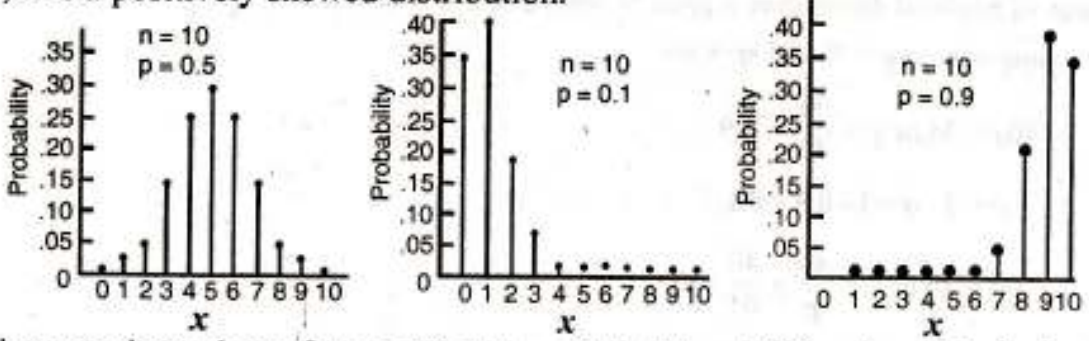
For (*i*), this follows from the fact that both n and p are positive and hence ${}^nC_x$, p^x, q^{n-x} all are positive. Consequently $f(x) \geq 0$.

For (*ii*), we know that binomial expansion of

$$(q + p)^n = q^n + {}^nC_1 q^{n-1} p + {}^nC_2 q^{n-2} p^2 + \ldots\ldots + {}^nC_x q^{n-x} p^x + \ldots\ldots + p^n$$

Therefore, $\Sigma f(x) = \Sigma {}^nC_x p^x q^{n-x} = (q + p)^n = (1 - p + p)^n = 1$.

The binomial distribution is a family of distributions since each different value of n or p specifies a different distribution. In this distribution n and p are called parameters. Regardless of the value of n, the distribution is symmetrical when $p = 0.5$. For small values of n, when p is greater than 0.5, the distribution is asymmetrical, with the peak occurring to the right of centre, *i.e.*, it is a negatively skewed distribution and when p is less than 0.5, the distribution is asymmetrical with the peak occurring to the left of the centre, *i.e.*, it is a positively skewed distribution.



The diagram given above for $n = 10$ and $p = 0.5$, 0.1 and 0.9 makes this distinction clear.

Mean and Variance of Binomial Distribution

The Mean. The mean of binomial random variable X, denoted by μ or $E(X)$, is the theoreti expected number of successes in n trials.

$$\mu = E(X) = \sum_{x=0}^{n} x\, f(x)$$

i.e., the mean of X is the sum of the products of the values that X can assume multiplied by their respect probabilities.

$$\mu = E(X) = \Sigma x f(x) = \Sigma x {}^{n}C_{x} p^{x} q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^{x} q^{n-x}$$

$$= \Sigma x \frac{n(n-1)!}{x(x-1)!(n-x)!} p^{x} q^{n-x}$$

$$= \sum_{x=1}^{n} np \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= np\Sigma^{n-1}C_{x-1} p^{n-1} q^{n-x} = np\,(q+p)^{n-1} = np \qquad [\because (q+p)^{n-1} =$$

Thus, the mean of the binomial distribution is np.

The Variance. The variance of the binomial random variable X measures the variation of the binon distribution and is given by

$$\sigma^2 = E(X^2) - \mu^2 = \sum_{x=0}^{n} x^2 f(x) - \mu^2$$

Here, $\mu = np$

and $\Sigma x^2 f(x) = \Sigma [x(x-1) + x]^{n} C_{x} p^{x} q^{n-x}$

$$= \Sigma x(x-1)^{n} C_{x} p^{x} q^{n-x} + \Sigma x^{n} C_{x} p^{x} q^{n-x}$$

$$= n(n-1) p^2 (q+p)^{n-2} + np$$

$$= n(n-1) p^2 + np \qquad [\because (q+p)^{n-2} =$$

Therefore, $\sigma^2 = n(n-1) p^2 + np - (np)^2$

$$= np[(n-1) p + 1 - np]$$

$$= np(1-p) = npq \qquad [\text{Since } p+q =$$

Thus, the standard deviation of the binomial distribution is $\sqrt{npq}$ and variance $= npq$.

Illustration 2. The mean of a binomial distribution is 40 and standard deviation 6. Calculate n, p and q.

Solution. The mean of binomial distribution is given by np and standard deviation by $\sqrt{npq}$.

Since $\sqrt{npq} = 6$; $npq = 36$ and $np = 40$

Therefore, $40q = 36$ or $q = \frac{36}{40} = 0.9$

$p = 1 - q = 1 - 0.9 = 0.1$

Given, $np = 40$ or $n = \frac{40}{p} = \frac{40}{0.1} = 400.$

Hence for the given question, $n = 400$, $p = 0.1$ and $q = 0.9$.

Illustration 3. Suppose that the half of the population of town are consumers of rice. One hundred investigators are appointed to find out its truth. Each investigator interviewed 10 individuals. How many investigators do you expect to report that three or less of the people interviewed are consumers of rice ? *(MBA, Bharathidasan Univ., Nov. 2001)*

Solution. Probability of a person being consumer of rice is $p = 1/2$, $q = 1/2$. Probability that three people or less are consumers of rice is given by

$$P[X \le 3] = P[X=0] + P[X=1] + P[X=2] + P[X=3]$$

$$= q^{10} + {}^{10}C_1\, q^9 p^1 + {}^{10}C_2\, q^8 p^2 + {}^{10}C_3\, q^7 p^3$$

$$= \left(\frac{1}{2}\right)^{10} + 10\left(\frac{1}{2}\right)^{10} + 45\left(\frac{1}{2}\right)^{10} + 120\left(\frac{1}{2}\right)^{10}$$

$$= \left(\frac{1}{2}\right)^{10}(1 + 10 + 45 + 120) = \frac{176}{1024}$$

Therefore, the number of investigators to report that three or less people are consumers of rice is given by $= \frac{176}{1024} \times 100 = 17.2 = 17$ approx.

Illustration 4. 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 4 or more will contract the disease ? *(MBA, DU, 2002, 2005)*

Solution. The probability of a worker who is suffering from the disease, *i.e.*, $p = \frac{20}{100} = \frac{1}{5}$

The probability of a worker who is not suffering from the disease

i.e., $$q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}$$

The probability of 4 or more, *i.e.*, 4, 5 or 6 will contract disease is given by

$$P[X \ge 4] = P[4] + P[5] + P[6]$$

$$= \left({}^6C_4\right)\left(\frac{1}{5}\right)^4 + \left({}^6C_5\right)\left(\frac{1}{5}\right)^5\left(\frac{4}{5}\right) + \left({}^6C_6\right)\left(\frac{1}{5}\right)^6$$

$$= 15\left(\frac{1}{5}\right)^4\left(\frac{4}{5}\right)^2 + 6\left(\frac{1}{5}\right)^5\left(\frac{4}{5}\right) + \left(\frac{1}{5}\right)^6$$

$$= \frac{15 \times 16}{15625} + \frac{6 \times 4}{15625} + \frac{1}{15625}$$

$$= \frac{1}{15625}[240 + 24 + 1] = \frac{265}{15625} = \frac{53}{3125} = 0.01696.$$

Illustration 5. Assume that on an average one telephone number out of fifteen is busy. What is the probability that if six randomly selected telephone numbers are called

(*a*) not more than three will be busy ?

(*b*) at least three of them will be busy ?

Solution. p = probability that a telephone number is busy $= \frac{1}{15}$

$$q = 1 - p = 1 - \frac{1}{15} = \frac{14}{15}; \text{ and } n = 6.$$

(*a*) The probability that out of six randomly selected telephone numbers not more than three numbers are busy is given by

$$P[X \le 3] = P(0) + P(1) + P(2) + P(3)$$

$$= \left({}^6C_0\right)\left(\frac{14}{15}\right)^6 + \left({}^6C_1\right)\left(\frac{14}{15}\right)^5\left(\frac{1}{15}\right)^1 + \left({}^6C_2\right)\left(\frac{14}{15}\right)^4\left(\frac{1}{15}\right)^2 + \left({}^6C_3\right)\left(\frac{14}{15}\right)^3\left(\frac{1}{15}\right)^3$$

$$= \left(\frac{1}{15}\right)^6\left[(14)^6 + 6(14)^5 + 15(14)^4 + 20(14)^3\right]$$

$$= \frac{(14)^3}{(15)^6}\left[(14)^3 + 6(14)^2 + 15(14) + 20\right]$$

$$= \frac{2744}{(15)^6}[2744 + 1176 + 210 + 20]$$

$$= \frac{2744 \times 4150}{11390625} = \frac{11387600}{11390625} = 0.9997$$

(*b*) Probability that at least three telephone numbers are busy is given by

$$P[X \geq 3] = P(3) + P(4) + P(5) + P(6)$$

$$= ({}^6C_3)\left(\frac{14}{15}\right)^3\left(\frac{1}{15}\right)^3 + ({}^6C_4)\left(\frac{14}{15}\right)^2\left(\frac{1}{15}\right)^4 + ({}^6C_5)\left(\frac{14}{15}\right)\left(\frac{1}{15}\right)^5 + ({}^6C_6)\left(\frac{1}{15}\right)^6$$

$$= 0.0051.$$

Fitting a Binomial Distribution. When a binomial distribution is to be fitted to observe data, t following procedure is adopted :

1. Determine the values of p and q. If one of these values is known, the other can be found out by t simple relationship $p = (1 - q)$ and $q = (1 - p)$. When p and q are equal, the distribution is symmetric Then p and q may be interchanged without altering the value of any term, consequently, terms equidist from the two ends of the series are equal. If p and q are unequal, the distribution is skewed. If p is l than 0.5, the distribution is positively skewed and when p is more than 0.5, the distribution is negativ skewed.

2. Expand the binomial $(q + p)^n$. The power n is equal to one less than the number of terms in t expanded binomial. Thus, when $n = 2$ there will be three terms in the binomial. Similarly, when n = there will be five terms.

3. Multiply each term of the expanded binomial by N (the total frequency), in order to obtain expected frequency in each category.

It is convenient to use the following recurrence relation for fitting of binomial distribution :

$$f(x) = P[X = x] = {}^nC_x p^x q^{n-x}$$

$$f(x+1) = P[X = x+1] = {}^nC_{x+1} p^{x+1} q^{n-x-1}$$

$$\frac{f(x+1)}{f(x)} = \frac{p}{q}\frac{n-x}{x+1}$$

$$f(x+1) = \frac{p}{q}\frac{n-x}{x+1}$$

When $x = 0$,

$$f(1) = \frac{p}{q}\frac{n-0}{0+1} f(0) = \frac{p}{q} n f(0)$$

When $x = 1$,

$$f(2) = \frac{p}{q}\frac{n-1}{2} f(1) = \left(\frac{p}{q}\right)^2 \frac{n(n-1)}{2!} f(0)$$

When $x = 2$,

$$f(3) = \frac{p}{q}\frac{n-1}{3} f(2) = \left(\frac{p}{q}\right)^3 \frac{n(n-1)(n-2)}{3!} f(0)$$

and so on.

This formula provides us a very convenient method for fitting the binomial distribution. The probability we need to calculate is $f(0)$ which is equal to q^n, where q can be estimated from the gi data.

Illustration 6. The screws produced by a certain machine were checked by examining number of defectives in a sample of 8. The following table shows the distribution of 128 samples according to the number of defective items they contained :

No. of defectives in a sample of 8 :	0	1	2	3	4	5	6	7	Total
No. of samples :	7	6	19	35	30	23	7	1	128

(*a*) Fit a binomial distribution and find the expected frequencies if the chance of machine being defective is ½.

(*b*) Find the mean and standard deviation of the fitted distribution. *(MBA, Delhi Univ., 2003)*

Solution. (*a*) The probability of a defective screw = ½.

$$p = ½, \quad q = ½, \quad N = 128.$$

Since there are 8 terms, therefore, $n = 7$.

The probability that 0, 1, 2, ..., 7 will be defective is given by expansion of :

$$(½ + ½)^7 = (½)^7 + ({}^7C_1)(½)^6(½)^1 + ({}^7C_2)(½)^5(½)^2 + ({}^7C_3)(½)^4(½)^3 + ({}^7C_4)(½)^3(½)^4 + ({}^7C_5)(½)^2(½)^5 + ({}^7C_6)(½)^6(½) + ({}^7C_7)(½)^7$$

$$= (½)^7 [1 + 7 + 21 + 35 + 35 + 21 + 7 + 1]$$

In order to obtain the expected frequencies, we shall multiply each term by N, *i.e.*, 128.

$$128\left(\frac{1}{2}+\frac{1}{2}\right)^7 = 128 \times \frac{1}{128}(1+7+21+35+35+21+7+1)$$

Thus, the expected frequencies are :

X :	0	1	2	3	4	5	6	7
f :	1	7	21	35	35	21	7	1

(*b*) Mean and standard deviation of the fitted distribution

Mean of the binomial distribution is np and standard deviation is $\sqrt{npq}$.

Here $n = 7, \quad p = ½, \quad q = ½$

Mean = np $\quad 7 \times ½ = 3.5$

$$\text{Standard deviation} = \sqrt{npq} = \sqrt{7 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{1.75} = 1.32.$$

Poisson Distribution

A second important discrete probability distribution is the Poisson distribution, named after the French mathematician S. Poisson who published its derivation in 1837.

The characteristics of the Poisson distribution are as follows :

1. The occurrence of the events is independent. That is, the occurrence of an event in an interval of space or time has no effect on the probability of a second occurrence of the event in the same, or any other interval.

2. Theoretically, an infinite number of occurrences of the event must be possible in the interval.

3. The probability of single occurrence of the event in a given interval is proportional to the length of the interval.

4. In any infinitesimal (extremely small) portion of interval, the probability of two or more occurrences of the event is negligible.

Poisson distribution differs from the binomial distribution in two important aspects :

(*a*) Rather than consisting of discrete trials, the distribution operates continuously over some given amount of time, distance, area, etc.

(*b*) Rather than producing a sequence of successes and failures, the distribution produces successes, which occur at random points in the specified time, distance, area. These successes are commonly referred to as 'occurrences'.

The Poisson distribution may be used to approximate binomial distribution when n is large and p is small and, therefore, is regarded as the limit of the binomial distribution.

The Poisson distribution is given by

$$f(x) = P(X = x) = \frac{e^{-m}m^x}{x!}; \quad x = 0, 1, 2.....$$

where, m is called the parameter of the distribution and is the average number of occurrences of random event, x is the number of occurrences of the random event and e is the constant whose value is 2.71

The Poisson distribution satisfies the two essential properties, *i.e.*, $f(x) \geq 0$ and $\Sigma f(x) = 1$.

The Poisson distribution has been extensively tabulated (see the Appendix). It has many applications in business and has been widely used in management science and operations research. The following are some of the examples which may be analysed with the use of this distribution :

(a) the demand for a product,

(b) typographical errors occurring on the pages of a book,

(c) the occurrence of accident in a factory,

(d) the arrival pattern in a departmental store,

(e) the occurrence of flaws in a bolt in a factory, and

(f) the arrival of calls at a switch board.

Mean and Variance of the Poisson Distribution

The Mean. The mean of the Poisson distribution is given by

$$\mu = E(X) = \Sigma x f(x) = \sum x \frac{e^{-m} m^x}{x!}$$

$$= 0 + me^{-m} + m^2 e^{-m} + \frac{m^3 e^{-m}}{2!} + \frac{m^4 e^{-m}}{3!} + \ldots\ldots$$

$$= me^{-m}\left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \ldots\ldots\right]$$

$$= me^{-m} e^{m} = m$$

Thus, the mean of the Poisson distribution is m.

The Variance. The variance of the Poisson distribution is given by

$$\sigma^2 = E(X^2) - [E(X)]^2 = E(X^2) - m^2$$

$$= \Sigma x^2 \frac{e^{-m} m^x}{x!} - m^2$$

But,
$$\Sigma x^2 \frac{e^{-m} m^x}{x!} = e^{-m} \Sigma \frac{[x(x-1) + x]}{x!} m^x$$

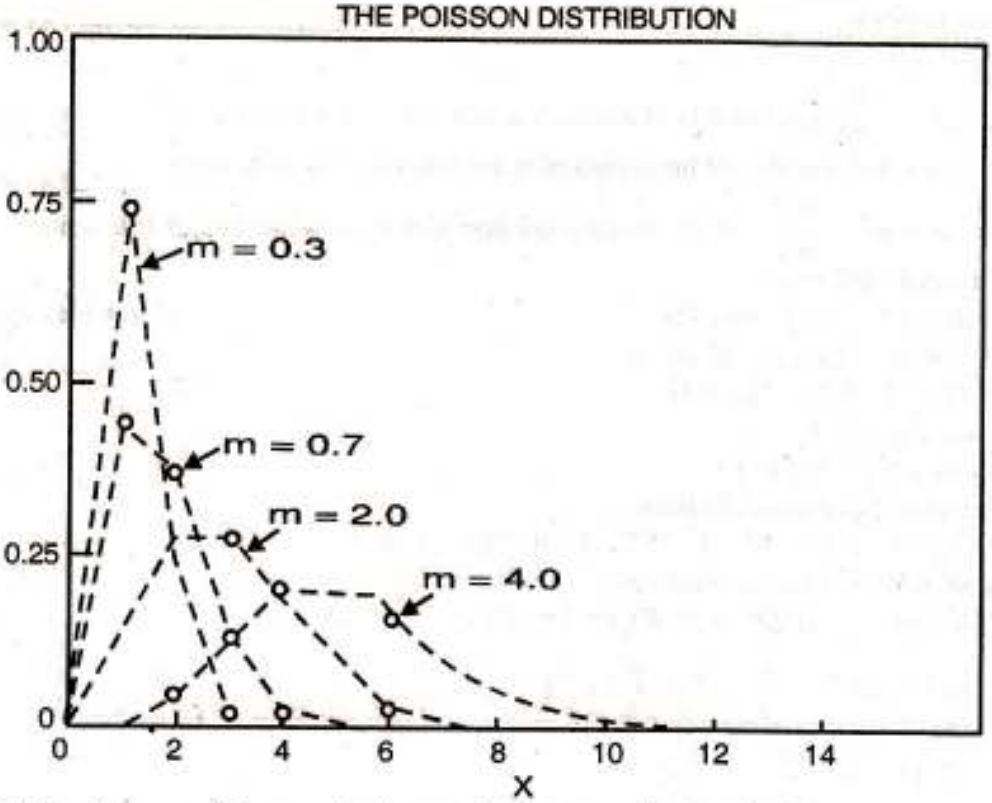
$$= m^2 e^{-m} \Sigma \frac{m^{x-2}}{(x-2)!} + me^{-m} \Sigma \frac{m^{x-1}}{(x-1)!} - m^2$$

$$= m^2 e^{-m} \cdot e^{m} + me^{-m} \cdot e^{m} - m^2$$

$$= m^2 + m - m^2 = m$$

Thus, the variance of the Poisson distribution is also equal to m.

The Poisson distribution is completely defined by the parameter m and is positively skewed. The positive skewness is typical of the Poisson distribution, indicating that, with extremely small probability there is the possibility that distribution will produce an indefinitely large number of occurrences in segment of time or space, even though the mean rate of occurrences may be quite small. As m increases the distribution shift to the right. This is illustrated on the next page, in the diagram for 4 values of from $m = 0.3$ to $m = 4.0$.



The Poisson distribution can frequently be used to approximate the binomial distribution when n is large and p is very small.

Form of the Poisson Distribution

Like binomial distribution, the variate of the Poisson distribution is also discrete one, *i.e.*, it takes only integral values. The probabilities of 0, 1, 2, ... occurrences are given by the successive terms of the expansion.

$$e^{-m}\left(1+m+\frac{m^2}{2!}+\frac{m^3}{3!}+\ldots\ldots+\frac{m^r}{r!}+\ldots\ldots\right)$$

This can be written in tabular form as follows :

No. of occurrences (x)	*Probability* $p(x)$	*No. of occurrences* (x)	*Probability* $p(x)$
0	e^{-m}	4	$\frac{m^4e^{-m}}{4!}$
1	me^{-m}	⋮	⋮
2	$\frac{m^2e^{-m}}{2!}$	⋮	⋮
3	$\frac{m^3e^{-m}}{3!}$	r	$\frac{m^re^{-m}}{r!}$
⋮	⋮	⋮	⋮

where $e = 2.7183$ and m is a constant called the parameter of the distribution, m is the average number of occurrences of an event.

The above table gives probabilities of 0, 1, 2, occurrences respectively. If we want to know the expected number of occurrences, we have to multiply each term by N, *i.e.*, the total number of observations.

Illustration 7. On the average, one in 400 items is defective. If the items are packed in boxes of 100, what is the probability that any given box of items will contain :

(*i*) no defectives ;

(*ii*) less than two defectives ;

(*iii*) one or more defectives ;
(*iv*) more than three defectives. *(MBA, Delhi Univ., 2005)*

Solution. Here, $p = \frac{1}{400}$; probability of a defective item which is very low.

$n = 100$; number of items packed in the box which is quite large

$m = np = \frac{100}{400} = 0.25$; average number of defectives in a box of 100 items

(*i*) Probability of no defective

$= P(X = 0) = e^{-m} = e^{-0.25} = 0.7788$ [From the table given in the appendix]

(*ii*) Probability of less than two defectives

$= P[X \le 1] = P[X = 0] + P[X = 1]$

$= e^{-m} + me^{-m} = e^{-m}(1 + m)$

$= 0.7788\ (1 + 0.25) = 0.9735$

(*iii*) Probability of one or more defectives

$= P[X \ge 1] = 1 - P[X = 0] = 1 - e^{-m} = 1 - 0.7788 = 0.2212$

(*iv*) Probability of more than three defectives

$= P[X \ge 4] = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$

$= 1 - [e^{-m} + me^{-m} + \frac{m^2}{2} e^{-m} + \frac{m^3}{6} e^{-m}]$

$= 1 - e^{-m}[1 + m + \frac{m^2}{2} + \frac{m^3}{6}]$

$= 1 - 0.7788[1 + 0.25 + 0.03125 + 0.0026]$

$= 1 - 0.7788[1.28385]$

$= 1 - 0.99986 = 0.00014$

Illustration 8. A factory produces blades in packets of 10. The probability of a blade to be defective is 0.2%. Find the number of packets having two defective blades in a consignment of 10,000 packets.

Solution. $m = np = 10 \times 0.002 = 0.02$

$$P[X = 2 \text{ defective blades}] = \frac{e^{-0.02}(0.02)^2}{2}$$

$$= \frac{0.9802 \times 0.0004}{2} = 0.4901 \times 0.0004 = 0.000196$$

Therefore, the total number of packets having two defective blades in a consignment of 10,000 packets is

$10{,}000 \times 0.000196 = 1.96$ or 2.

Illustration 9. What probability model is appropriate to describe a situation where 100 misprints are distributed randomly throughout the 100 pages of a book ? For this model, what is the probability that a page observed at random will contain at least three misprints ?

Solution. Since there are 100 misprints in 100 pages, it implies that there is only one mistake on the average in a page. Therefore, the probability of being a misprint is very small as a page contains large number of words and n the number of words in 100 pages will be very large. So, in this case probability of being a misprint is small and n is very large, therefore, Poisson distribution is best suited here.

Average or expected number of misprints in one page is

$$m = np = 100 \times 0.01 = 1$$

$$e^{-m} = e^{-1} = 0.3679$$

Probability of at least three misprints in a page is

$$= P[X \ge 3] = 1 - P(X < 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

$$= 1 - [e^{-1} + e^{-1} + \frac{e^{-1}}{2\,!}]$$

$$= 1 - e^{-1}[2.5] = 1 - 0.3679\ (2.5)$$

$$= 1 - 0.9198 = 0.0802.$$

Fitting a Poisson Distribution. The process of fitting a Poisson distribution is simple. We have to obtain the values of m and calculate the probability of zero occurrences. The other probabilities can be easily calculated by the recurrence relation as follows :

$$f(x) = \frac{e^{-m}m^x}{x!}, \text{ and } f(x+1) = \frac{e^{-m}m^{x+1}}{(x+1)!}$$

$$\frac{f(x+1)}{f(x)} = \frac{m}{x+1} \text{ or } f(x+1) = \frac{m}{x+1} f(x)$$

when $x = 0$, $f(1) = m f(0)$

when $x = 1$, $f(2) = \frac{m}{2} f(1) = \frac{m^2}{2} f(0)$

when $x = 2$, $f(3) = \frac{m}{3} f(2) = \frac{m^3}{6} f(0)$

and so on.

This recurrence relation provides a very convenient method for fitting the Poisson distribution. The only probability we need to know is $f(0) = e^{-m}$, where m is the only parameter of the Poisson distribution.

Multiplying by N (the total frequency) each probability of the Poisson distribution, we get the expected frequencies for respective probabilities.

Illustration 10. The following table gives the number of days in a 50-day period during which automobile accidents occurred in a city. Fit Poisson distribution to the data :

No. of accidents :	0	1	2	3	4
No. of days :	21	18	7	3	1

(MBA, Kumaun Univ., 2006)

Solution. FITTING OF POISSON DISTRIBUTION

X	f	fX
0	21	0
1	18	18
2	7	14
3	3	9
4	1	4
	$N = 50$	$\Sigma fX = 45$

$$m = \bar{X} = \frac{\Sigma fX}{N} = \frac{45}{50} = 0.9$$

$f(0) = e^{-m} = e^{-0.9} = 0.4066$

$f(1) = m f(0) = (0.9)(0.4066) = 0.3659$

$f(2) = \frac{m}{2} f(1) = \frac{0.9}{2} (0.3659) = 0.1647$

$f(3) = \frac{m}{3} f(2) = \frac{0.9}{3} (0.1647) = 0.0494$

$f(4) = \frac{m}{4} f(3) = \frac{0.9}{4} (0.0494) = 0.0111$

In order to fit Poisson distribution, we shall multiply each probability by N, *i.e.*,50.

Hence, the expected frequencies are :

X:	0	1	2	3	4
f:	0.4066×50	0.3659×50	0.1647×50	0.0494×50	0.0111×50
	$= 20.33$	$= 18.30$	$= 8.24$	$= 2.47$	$= 0.56$

Negative Binomial Distribution

Where as the binomial distribution describes the probabilities of the number of successes likely to appear in a sequence of *fixed* number of Bernoulli trials, the negative binomial distribution,

as the name itself implies, describes the probabilities of the number of trials likely to be required in order to obtain a *fixed* number of successes.

To derive the probability mass function of the negative binomial distribution, we proceed as follows: suppose that x trials are required to obtain exactly k successes ($x \geq k$; *i.e.*, $x = k, k+1, k+2, ...$). Then, clearly, the xth trial should yield the kth success, the previous $(x-1)$ trials should give the remaining $(k-1)$ successes and $(x-1)-(k-1) = (x-k)$ failures in some sequence of successes and failures.

Clearly, this can happen in $\binom{x-1}{k-1}$ distinct ways. Now, observe that each particular sequence has k successes and $(x-k)$ failures in itself and hence its probability of occurrence is $p^k(1-p)^{x-k}$. As the sequence of trials are Bernoullian, the probability that exactly x trials will be required to obtain k successes becomes

$$\binom{x-1}{k-1} p^k (1-p)^{x-k}; \qquad x = k, k+1, \ldots$$

Thus, if x is the random variable corresponding to the number of trials required for observing exactly k successes in an indefinite sequence of Bernoullian trials, its distribution is said to be negative binomial and has the following probability mass function.

$$P[x\,;k,p] = \binom{x-1}{k-1} p^k (1-p)^{x-k} \qquad \text{if } x = k, k+1, \ldots$$

$$= 0, \text{ otherwise}$$

where $0 < p < 1$ and $k \geq 1$ is a fixed integer.

The negative binomial distribution arises in practice, where observation of successes takes place as a waiting-time phenomenon.

For the negative binomial variate, it is easy to prove that :

$$\text{Mean} = \frac{k}{p} \text{ and variance} = \frac{k(1-p)}{p^2}$$

Illustration 11. A market research agency that conducts interviews by telephone has found, from past experience, that there is a 0.40 probability that a call made between 2.30 and 5.30 P.M. will be answered. Assuming a Bernoullian process :

(*a*) What is the probability that an interviewer's tenth answer comes on his twentieth call?

(*b*) What is the expected number of calls necessary to obtain seven answers ?

(*c*) What is the probability that an interviewer will receive his first answer on his third call?

Solution. Let success denote an answer to a call. Then $p = 0.40$.

(*a*) Prob. (10th answer comes on 20th call) $= \binom{x-1}{k-1} p^k (1-p)^{x-k}$

where $x = 20$ and $k = 10$

$$= \binom{20-1}{10-1} (0.4)^{10} (0.6)^{10} = 0.05856$$

(*b*) Expected number of calls for seven answers $= \frac{k}{p}$

where $k = 7$ and $p = 0.4$

$$= \frac{7}{0.4} = \frac{70}{4} = 17.5$$

(*c*) Prob. (1st answer on 3rd call) $= \binom{3-1}{1-1} (0.4)^2 (0.6)^2 = 0.1440.$

Multinomial Distribution

The multinomial distribution is the multivariate analogue of the binomial distribution. It is one of the simplest but most important discrete multivariate distributions. The binomial distribution arises from a random experiment in which a finite sequence of n repeated and independent Bernoulli trials are conducted, each trial resulting in only one or two possible outcomes, success or failure. The multinomial distribution arises when the number of possible outcomes of a single trial is generalised from 2 to k.

Let $E_1, E_2, \ldots\ldots, E_k$ denote the k possible outcomes in a single trial of the experiment. Let n repeated trials be conducted and the various outcomes noted. Let X_i denote the number of times the outcome E_i is observed with corresponding probabilities p_i $(i = 1, 2, \ldots, k)$. As the outcomes are mutually exclusive,

$$X_1 + X_2 + \cdots\cdots + X_k = n$$

and $$p_1 + p_2 + \cdots\cdots + p_k = 1$$

Then, the joint probability mass function of $X_1, X_2, \ldots\ldots, X_k$ is called the multinomial distribution and is given by

$$P[X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k] = \frac{n!}{x_1!\, x_2! \ldots x_k!}\, p_1^{x_1} p_2^{x_2} \ldots\ldots p_k^{x_k}$$

where $X_1, X_2, \ldots\ldots, X_k$ are non-negative integers such that

$$\sum_{i=1}^{k} X_i = n$$

Illustration 12. Consider the production of ball bearings of a certain type, the diameter of which should be 0.2500 inch. Because of the inherent variability in the manufacturing process, the bearings are classified as 'undersize' 'oversize' and 'acceptable' if they measure less than 0.2495 inch, more than 0.2505 inch, and between 0.2495 and 0.2505 inch, respectively. Suppose the production process for three bearings is such that 4% of the bearings are undersize, 6% are oversize, and 90% are acceptable. If 100 of these bearings are picked at random, what is the probability of getting x_1 undersize, x_2 oversize, and x_3 acceptable bearings?

Solution. Here, $n = 100$, $p_1 = 0.04$, $p_2 = 0.06$ and $p_3 = 0.90$

The required probability is given by

$$P(x_1, x_2, x_3) = \frac{100!}{x_1! x_2! x_3!}(0.04)^{x_1}(0.06)^{x_2}(0.90)^{x_3}$$

where $0 \le x_i \le 100; i = 1, 2, 3$

and $$\sum_{i=1}^{k} X_i = 100.$$

Illustration 13. In the above illustration, suppose 6 bearings are sampled from the process.

(*i*) What is the probability that there will be two bearings of each category?

(*ii*) What is the probability that all of them will be accepted?

Solution. (*i*) Here, $n = 6$, $x_1 = 2$, $x_2 = 2$, and $x_3 = 2$

The required probability is given by

$$= \frac{6!}{2!\,2!\,2!}(0.04)^2 (0.06)^2 (0.9)^2$$

$$= 90\,(0.00000576)\,(0.81) = 0.0004199$$

(*ii*) Here, $n = 6$, $x_1 = 0$, $x_2 = 0$, and $x_3 = 6$

$$\text{Prob. [all will be acceptable]} = \frac{6}{0!\,0!\,6!}(0.04)^0 (0.06)^6 (0.90)^6 = 0.5314.$$

Hypergeometric Distribution

The Hypergeometric distribution arises when a simple random sample without replacement is drawn from a dichotomous population (*i.e.*, one whose elements can be divided into two mutually exclusive categories) consisting of finite number of elements in each of the two categories and the random variable observed is the number of items of one particular category present in the sample. Suppose a box contains numbers of a manufactured product, of which n_1 are good and $n - n_1$ are defective. Suppose a sample of size R is drawn, then the number X of good items in the sample is a random variable taking values 0, 1, 2....., K with probabilities,

$$P(X) = \frac{\binom{n_1}{X}\binom{n-n_1}{K-X}}{\binom{n}{K}} \qquad \text{if } X = 0,1,.....K$$

$$= 0, \text{ otherwise.}$$

This probability mass function is called hypergeometric distribution.

It is easy to verify that the mean and the variance of the above distribution are given by :

$$\text{Mean} = \frac{n_1}{n} \text{ and Variance} = K\frac{n_2}{n}\frac{n-n_1}{n}\frac{n-K}{n-1}$$

Illustration 14. Certain missile components are shipped in lots of 12. Three components are selected from each lot and a particular lot is accepted if none of the three components selected is defective.

(*a*) What is the probability that a lot will be accepted if it contains 5 defectives?

(*b*) What is the probability that a lot will be rejected if it contains 9 defectives?

(*c*) Let X be a random variable denoting the number of defectives in a sample of 3 components selected randomly from one of the above lots. If the lot contains 4 defectives, specify the probability function $f(x)$. Present the probability distribution (*i*) as a mathematical expression, and (*ii*) in the form of a table.

(*d*) Under the conditions stated in (*c*) above, what is the expected number of defectives in a sample of 3 components?

Solution. (*a*) Here, $n = 12$, $n_1 = 5$.

The lot will be accepted if the 3 components randomly selected from this lot has no defectives. The required probability is, therefore,

$$\frac{\binom{7}{3}\binom{5}{0}}{\binom{12}{5}} = \frac{7}{44} = 0.1591$$

(*b*) Here, $n = 12$, $n_1 = 9$.

The lot will be rejected if at least one of the components randomly chosen from this lot is defective.

This probability is given by

$$1 - \text{Prob [none is defective]} = 1 - \frac{\binom{3}{3}\binom{9}{0}}{\binom{12}{3}} = 1 - \frac{1}{220} = 1 - 0.0045 = 0.9955$$

(*c*) The lot contains 4 defectives, therefore,

Number of defectives = 4

Number of non-defectives = 8

X = random variable denoting the number of defectives in a sample of 3 components from the above lot.

probability function is

$$f(x) = \frac{\binom{4}{X}\binom{8}{3-X}}{\binom{12}{3}}; X = 0, 1, 2, 3$$

$= 0$, otherwise

table of probabilities is

X	$f(x)$
0	0.2545
1	0.5091
2	0.2182
3	0.0182
	$\Sigma f(x) = 1$

The expected number of defectives in a sample of 3 components

$= K.\frac{n_1}{n}$ where $K = 3$, $n_1 = 4$ and $n = 12$

$= \frac{3 \times 4}{12} = 1.$

Normal Distribution

The Normal* distribution was discovered by De Moivre as the limiting case of Binomial model in 1733. It was also known to Laplace no later than 1774, but through a historical error it has been credited to Gauss, who first made reference to it in 1809. Throughout the 18th and 19th 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 failed because of the false premises. The normal model has, nevertheless, become the most important probability model in statistical analysis.

The normal distribution is approximation to binomial distribution. Whether or not p is equal to q, the binomial distribution tends to the form of the continuous curve when n becomes large at least for the central part of the range. As a matter of fact, that correspondence between binomial and the normal curve is surprisingly close even for low values of n provided p and q are fairly near equality. The limiting frequency curve, obtained as n, becomes large and is called the normal frequency curve or simply the normal curve.

A random variable X is said to have a normal distribution with parameters μ (mean) and σ^2 (variance) if the density function is given by :

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \qquad -\infty < X < \infty$$

where y = the computed height of an ordinate at a distance of X from the mean,

σ = Standard deviation of the given normal distribution,

π = the constant = 3.1416 ; $\sqrt{2\pi}$ = 2.5066,

e = the constant = 2.7183 (the base of the system of natural logarithm),

μ = Mean of the given normal distribution.

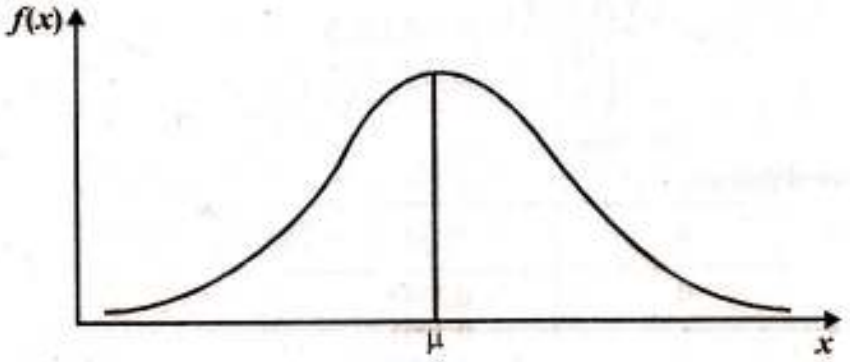
In symbols, it can be expressed as

$$X \sim N(\mu, \sigma)$$

This is read as : The random variable X follows normal distribution with mean μ and standard deviation σ.

*The word normal does not simply mean that the other distributions are abnormal. It is simply the name customarily given to this distribution. Sometimes, the name *Gaussian* is used instead of *normal* to avoid confusion.

If we draw the graph of normal distribution, the curve obtained will be known as normal curve and is given below.



The graph of $y = f(x)$ 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.

When we say that curve has unit area, we mean that the total frequency N is equated to 1. To obtain ordinates for particular distribution the ordinates given by the above formula are multiplied by N. The equation to a normal curve corresponding to a particular distribution is given by

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

The quantity $\frac{N}{\sigma\sqrt{2\pi}}$ in the above formula is equal to the maximum ordinate of the normal curve which will always occur at the mean of the distribution.

A random variable with any mean and standard deviation can be transformed to a standardized normal variable by subtracting the mean and dividing by the standard deviation. For a normal distribution with mean μ and standard deviation σ, the standardized variable z is obtained as

$$z = \frac{X - \mu}{\sigma}$$

A value z represents the distance, expressed as a multiple of the standard deviation, that the value X lies away from the mean. The standardized variable z is called a *standard normal variate* which has mean zero and standard deviation one. In symbols, if

$$X \sim N(\mu, \sigma)$$

then

$$z \sim N(0, 1)$$

The probability density function of the standard normal variate z is given by

$$y = f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad -\infty < z < \infty$$

Relation between Binomial, Poisson and Normal Distribution

The three distributions, namely Binomial, Poisson and Normal, are very closely related to each other. As explained earlier, when n is large and the probability p of occurrence of an event is close to zero so that np remains a finite constant, then the binomial distribution tends to Poisson distribution.

Similarly, there is a relation between binomial and normal distributions. Normal distribution is a limiting form of binomial distribution under the following conditions :

(1) n, the number of trials is very large, *i.e.*, $n \to \infty$, and

(2) neither p nor q is very small.

In fact, it can be proved that the binomial distribution approaches a normal distribution with standardized normal variable, *i.e.*,

$$z = \frac{X - np}{\sqrt{npq}} \sim N(0, 1)$$

or, $\frac{X - np}{\sqrt{npq}}$ will follow the normal distribution with mean zero and variance one.

Similarly, Poisson distribution also approaches a normal distribution with standardized normal variable, *i.e.*,

$$z = \frac{X - m}{\sqrt{m}} \sim N(0, 1)$$

In other words, $\frac{X - m}{\sqrt{m}}$ will follow the normal distribution with mean zero and variance one.

The Standard Deviation and the Normal Curve

In any normal curve, an exact percentage of observations in the distribution falls within ranges established by the standard deviation in conjunction with the mean. If we cut through the distribution at one, two, and three standard deviation away from mean, on both sides, we obtain the areas of the distribution which contain certain percentages of all the observations.

In a normal curve, between the range of arithmetic mean plus 1 standard deviation and minus 1 standard deviation, *i.e.*, $\mu \pm 1\sigma$ covers 68.27% of the observations in the distribution and 34.13% of observations will fall on either side of mean. Similarly, $\mu \pm 2\sigma$ covers 95.45% of observations and $\mu \pm 3\sigma$ covers 99.73% of observations.

The table shows the area of the normal curve between mean ordinate and ordinates at various sigma (σ) distances from the mean (μ) as percentage of the total area.

AREA RELATIONSHIP

Distance from the mean ordinate	*Percentage of total area*
0.5 σ	19.146
1.0 σ	34.135
1.5 σ	43.319
1.96 σ	47.500
2.00 σ	47.725
2.5 σ	49.379
2.5758 σ	49.500
3.0 σ	49.865

Thus, the two ordinates at distance 1.96 σ from the mean on either side would enclose 47.5 + 47.5 = 95% of the total area and two ordinates at 2.5758 σ distance from the mean on either side would enclose 49.5 + 49.5 = 99% of the total area. The area enclosed between ordinates at 3σ distance from the mean on either side would be 49.865 + 49.865 = 99.73% of the total area.

Moments of the Normal Distribution

For the normal distribution, all odd order moments about mean are zero and given by the relation

$$\mu_1 = \mu_3 = \mu_5 = \ldots\ldots = \mu_{2n+1} = 0$$

and the even order moments about mean is given by the relation

$$\mu_{2n} = 1.3.5 \ldots\ldots\ldots\ldots (2n-1)\sigma^{2n}$$

In particular, we have

$$\mu_2 = 1.\sigma^2 = \sigma^2; \mu_4 = 1.3.\sigma^4 = 3\sigma^4;$$

Therefore, moment coefficient of skewness β_1 is given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \qquad [\because \mu_3 =$$

and moment coefficient of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$$

Hence, $\beta_1 = 0$ shows that normal distribution is perfectly symmetrical and $\beta_2 = 3$ indicates that the normal curve is mesokurtic in shape. If $\beta_2 < 3$, the curve is platykurtic and if $\beta_2 > 3$, the curve is leptokurtic.

Properties of the Normal Distribution

The following are the important properties of the normal curve and the normal distribution:

1. The normal curve is symmetrical about the mean (skewness = 0). If the curve is folded along its vertical axis, the two halves will coincide. The number of cases below the mean in a normal distribution is equal to the number of cases above the mean, which makes the mean and median coincide. The height of the curve for a positive deviation of 3 units is the same as the height of the curve for a negative deviation of 3 units.

2. The height of the normal curve is at its maximum at the mean. Hence, the mean and mode of the normal distribution coincide. Thus for a normal distribution mean, median and mode are all equal.

3. There is one maximum point of the normal curve which occurs at the mean. The height of the curve declines as we go in either direction from the mean. The curve approaches nearer and nearer to the base but it never touches it. In other words, the curve is asymptotic to the base on either direction. Hence, its range is unlimited or infinite on both directions.

4. Since there is only one maximum point, therefore, the normal curve is unimodal, *i.e.*, it has only one mode.

5. The points of inflection occur at

$$x = \mu \pm \sigma, y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}}$$

6. As distinguished from binomial and Poisson distributions where variable is discrete, the variable distributed according to the normal curve is a continuous one.

7. First and third quartiles are equidistant from the median.

8. Mean deviation about mean = $\frac{4}{5}\sigma$ or, more precisely, 0.7979 times of standard deviation.

9. Linear combination of independent normal variates is also a normal variate, *i.e.*, if X_1 and X_2 are two independent normal variates and a_1 and a_2 are given constants, then the linear combination $a_1X_1 + a_2X_2$ will also follow a normal distribution.

10. All odd moments of the normal distribution are zero.

i.e., $\mu_{2n+1} = 0$ $\qquad (n = 0, 1, 2, \ldots$

11. $\beta_1 = 0$ and $\beta_2 = 3$

since $\beta_1 = 0$, therefore, the normal distribution is perfectly symmetrical and $\beta_2 = 3$ implies that normal curve is neither leptokurtic nor platykurtic.

12. Mean $\pm \sigma$, mean $\pm 2\sigma$, and mean $\pm 3\sigma$ covers 68.27%, 95.45%, and 99.73% area respectively.

Importance of Normal Distribution

The normal distribution has great significance in statistical work because of the following reasons :

1. The normal distribution has the remarkable property stated in the so-called central limit theorem*, which asserts that certain statistics, most important of which is the sample mean and sample variance, tends to be normally distributed as the sample size becomes large.

2. Even if a variable is not normally distributed, it can sometimes be brought to normal form by simple transformation of variable. For example, if distribution of X is skewed, the distribution of $\sqrt{X}$ might come out to be normal.

3. Many of the sampling distributions like Student's t, F, etc., also tend to normal distribution.

4. The sampling distribution and tests of hypothesis are based upon the assumption that samples have been drawn from a normal population with mean μ and variance σ^2.

5. Normal distribution find large applications in Statistical Quality Control.

6. As n becomes large, the normal distribution serves as a good approximation for many discrete distributions (such as Binomial, Poisson, etc.).

7. In theoretical statistics, many problems can be solved only under the assumption of a normal population. In applied work, we often find that methods developed under the normal probability law yield satisfactory results, even when the assumption of a normal population is not fully met, despite the fact that the problem can have a formal solution only if such a premise is hypothesized.

8. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate. For example, the moments of the normal distribution are expressed in simple form. The normal curve is reasonably close to many distributions of the humped type. If, therefore, we are ignorant of the exact nature of a humped distribution, or know the form but find it mathematically interactive, we may assume as a first approximation that the distribution is normal and see where this assumption leads us.

The admiration of normal distribution has been beautifully expressed by a well-known statistician W.J. Youden in the following words :

> THE
> NORMAL
> LAW OF ERROR
> STANDS OUT IN
> THE EXPERIENCE OF
> MANKIND AS ONE OF THE
> BROADEST GENERALISATION OF
> NATURAL PHILOSOPHY. IT SERVES AS THE
> GUIDING INSTRUMENT 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.

Artistically enough, it gives us the shape of the normal curve also.

Area under the Normal Curve

Since for different values of μ and σ, we shall have different normal curves, therefore, it may be very difficult to find areas under the normal curves for different pair of values of μ and σ. Hence, the areas for a normal curve are tabulated in terms of the standardized normal variate z. As any normally distributed random variate X with parameter μ and σ can be transformed to the standardized normally distributed random variate z, therefore, the table given in the appendix under the heading of area under the normal curve may be used.

*See Chapter on Sampling and Sampling Distributions.

This table in the appendix contains the probabilities for the area under the normal curve between mean $z = 0$ and any other specified value of z. As the normal curve is symmetrical, therefore, the area under the normal curve is given only for half of the positive side of the curve.

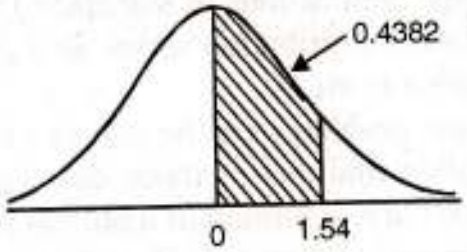
For example, corresponding to $z = 1$, the area under the normal curve is given as 0.3413 (from the table in the appendix). Therefore, for $z = -1$, the area is also 0.3413.

Hence, $Pr\ [-1 \leq z \leq +1] = 0.3413 + 0.3413 = 0.6826$.

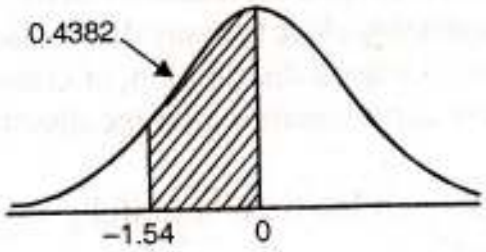
The following are some of the examples to illustrate how tables are to be used in order to obtain area under the normal curve.

Illustration 15. Find out the area under the normal curve for $z = 1.54$.

Solution. If we look to the table, the entry corresponding to $z = 1.54$ is 0.4382 and this gives the area in the shaded region in following figure between $z = 0$ and $z = 1.54$.



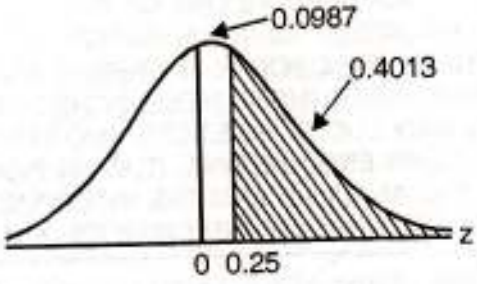
The table given at the end of the book does not contain entries corresponding to negative values of Z. But, since the curve is symmetrical, we can find the area between $z = 0$ and $z = -1.54$ by looking up the area corresponding to $Z = 1.54$.



If we wish to cut the area under normal curve to the right of a positive value of z, we should subtract the tabular value from 0.5000. The reason is that the normal curve is symmetrical, the area to the right of the mean is 0.5000 and the area to the right of a positive value of z is 0.5000 minus the tabular value given for z.

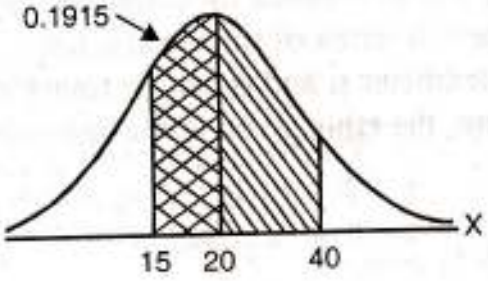
Illustration 16. Find the area to the right of $z = 0.25$.

Solution. Subtract 0.0987 (the entry given in the table for $z = 0.25$) from 0.5000 getting $(0.5000 - 0.0987) = 0.4013$ as shown below :



If we wish to find out the area to the left of a positive value of z, we add 0.5000 to the tabular value given for z.

Illustration 17. A normal curve has $\mu = 20$ and $\sigma = 10$. Find the area between $X_1 = 15$ and $X_2 = 40$.



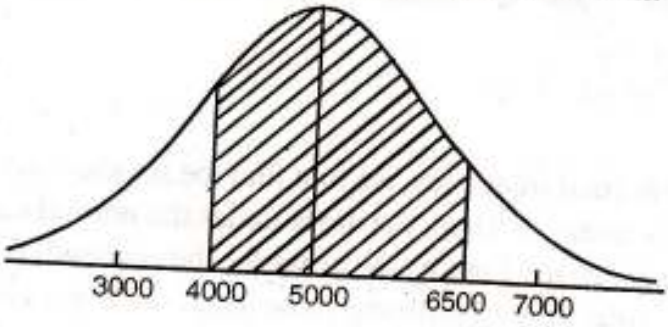
Solution. $z_1 = \dfrac{X-\mu}{\sigma} = \dfrac{15-20}{10} = -0.5$

$$z_2 = \frac{40-20}{10} = +2.0.$$

Consulting the table, we find the areas corresponding to the z's are 0.1915 and 0.4772 and thus the desired area between $X_1 = 15$ and $X_2 = 40$ is $(0.1915 + 0.4772) = 0.6687$.

Illustration 18. How many workers have a salary between Rs. 4000 and Rs. 6500, if the arithmetic mean is Rs. 5000, standard deviation is Rs. 1000 and number of worker is 15,000, if the salary of the worker is assumed to follow the normal law ?

Solution.



$$z_1 = \frac{4000-5000}{1000} = -1 \qquad \text{(left of the mean)}$$

$$z_2 = \frac{6500-5000}{1000} = 1.5 \qquad \text{(right of the mean)}$$

From the table, we find that 34.13% of workers fall between Rs. 4000 and Rs. 5000 and 43.32% fall between Rs. 5000 and Rs. 6500.

$\therefore$ 34.13 + 43.32 = 77.45% of workers have a salary between Rs. 4000 and Rs. 6500.

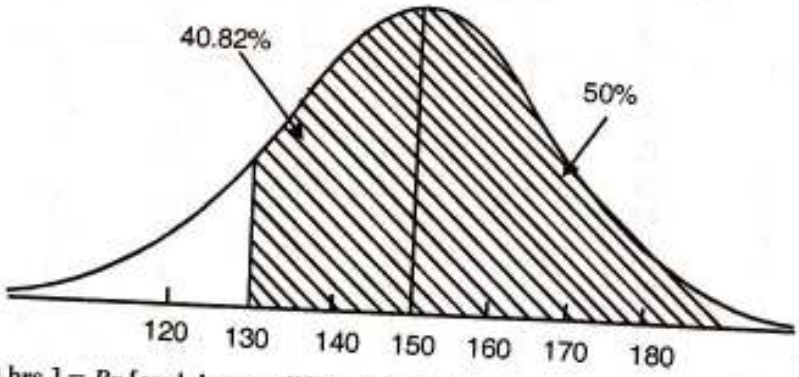
$\therefore$ Number of workers getting a salary between Rs. 4000 and Rs. 6500 is given by

$$0.7745 \times 15{,}000 = 11{,}618$$

Illustration 19. A large flashlight is powered by 5 batteries. Suppose that the life of a battery is normally distributed with $\mu = 150$ hours and $\sigma = 15$ hours. The flashlight will cease functioning if one or more of its batteries go dead. Assuming the lives of batteries are independent, what is the probability that flashlight will operate more than 130 hours ?

Solution. $z = \dfrac{X-\mu}{\sigma} = \dfrac{130-150}{15} = -1.33$

From the table, we find the 40.82% of the batteries will operate more than 130 hours.



Pr [Flashlight life > 130 hrs.] = Pr [each battery life > 130 hrs.]

Since the life of batteries are independent

= Pr [each battery life > 130 hrs.]^5 = $[0.9082]^5$

Hence, the probability that the flashlight will operate more than 130 hours is given by $(0.9082)^5$.

Applications of the Normal Distribution

The normal distribution is mostly used for the following purposes :

1. To approximate or "fit" a distribution of measurement under certain conditions.

2. To approximate the binomial distribution and other discrete or continuous probability distributions under suitable conditions.

3. To approximate the distributions of means and certain other statistic calculated from samples, especially large samples.

Fitting of Normal Distribution

In order to fit a normal distribution to the given data, we first calculate the mean μ and standard deviation σ from the given data. Then the normal curve fitted to the given data is given by

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \qquad -\infty < X < \infty$$

To calculate the expected normal frequency, we first find the standard normal variates corresponding to the lower units of each class-interval. Then, the areas under the normal curve are computed from the tables for each standard normal variate. Finally, the areas for the successive class-intervals are obtained by subtracting the successive areas and multiplying these areas by N, we get the normal frequencies.

The following illustration shall illustrate the applications of normal distribution :

Illustration 20. The following table gives the distribution of height stature (in cms) among the management trainees in Delhi :

Height (in cms)	*No. of trainees*	*Height (in cms)*	*No. of trainees*
161	2	168	126
162	10	169	109
163	11	170	87
164	38	171	75
165	57	172	23
166	93	173	9
167	106	174	4

Test the normality of the distribution by comparing the proportion of cases lying between $\bar{X} \pm 1\sigma$, $\bar{X} \pm 2\sigma$, $\bar{X} \pm 3\sigma$ for the above distribution.

Solution. CALCULATION OF $\bar{X}$ AND σ

X	f	d	fd	fd^2
61	2	–6	–12	72
62	10	–5	–50	250
63	11	–4	–44	176
64	38	–3	–114	342
65	57	–2	–114	228
66	93	–1	–93	93
67	106	0	0	0
68	126	+1	+126	126
69	109	+2	+218	436
70	87	+3	+261	783
71	75	+4	+300	1,200
72	23	+5	+115	575
73	9	+6	+54	324
74	4	+7	+28	196
	N =750		Σfd = 675	Σfd^2 = 4801

$$\bar{X} = A + \frac{\Sigma fd}{N} = 167 + \frac{675}{750} = 167.9$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{4801}{750} - \left(\frac{675}{750}\right)^2}$$

$$= \sqrt{6.40 - 0.81} = \sqrt{5.59} = 2.36$$

$\bar{X} \pm 1\sigma = 167.9 \pm 2.36 = 165.54$ and 170.26

Number of trainees having stature in this range

$= 93 + 106 + 126 + 109 + 87 = 521.$

Therefore, Proportion $= \frac{521}{750} = 0.69$ or 69%

$\bar{X} \pm 2\sigma = 167.9 \pm 2\,(2.36) = 167.9 \pm 4.72 = 163.18$ and 172.62

Number of trainees having stature in this range

$= 11+38+57+93+106+126+109+87+75+23=725.$

Therefore, Proportion $= \frac{725}{750} = 0.96$ or 96%

$\bar{X} \pm 3\sigma = 167.9 \pm 3\,(2.36) = 167.9 \pm 7.08 = 160.82$ and 174.98

Number of trainees having stature in this range = 750 and hence 100%. In a normal distribution, the proportion lying between these limits is about 68%, 95%, and 99% respectively. Hence, the given distribution is approximately normal.

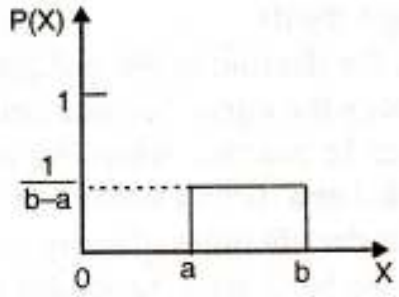
Uniform Distribution

A continuous random variable X is said to have a uniform distribution with parameters a and b if its probability density function is given by

$$P(x) = \frac{1}{b-a}, \qquad \text{if } a \le x \le b$$

$$= 0, \text{ otherwise}$$

The uniform distribution is also known as the 'Constant Probability Distribution', as the probability is constant [equal to $1/(b-a)$] at every point of the continuum (a, b) and is independent of whatever value the variable may take within the interval. Yet another name for the uniform distribution is 'rectangular distribution', for the shape of the curve of the probability density function is rectangular, as shown below :



It is easy to check that $P(x)$ defined as above is indeed a probability density function.

The mean and variance of a uniform distribution are given by

$$\text{Mean} = \frac{b+a}{2}; \quad \text{Variance} = \frac{(b-a)^2}{12}$$

The mean, variance and, in fact, all higher order moments of the uniform distribution depend solely on the values of a and b, the lower and upper bounds of the range of values the variable takes. Another interesting property of the uniform distribution is that events described by sub-intervals [of the main interval (a, b)] of equal length have equal probabilities of occurrence. This is a direct consequence, of the *constant* distribution of the total probability over the entire range.

The uniform distribution arises in practice whenever the probability of occurrence of the eve under consideration is constant whatever be the value of the variable, *i.e.*, all possible values of t continuous variable are assumed equally likely. For instance, commuter travel time between specifi points has been considered as a random variable with constant probabilities over a small range of tim

Exponential Distribution

A continuous random variable x is said to have an exponential distribution with parameter λ if i probability density function is given by

$$P(x) = \lambda e^{-\lambda x} \qquad \text{if } 0 \le x \le \infty, \lambda >$$
$$= 0, \text{ otherwise.}$$

At the outset, we shall note that whereas the uniform (or rectangular) variable takes values over *finite* range, the exponential variable takes values over an *infinite* range. It is easy to verify that th exponential density function defined as above is indeed a probability density function. We note that th only condition on λ is that it should be a positive real number. Hence by giving different values fo λ, different exponential distributions can be specified. This would help us to understand the natur of the exponential distribution better. The following table gives the ordinates of the exponenti probability density function for $x = 0$ to 6 for $\lambda = 0.2, 0.5, 1$ and 2.

x \ λ	0.2	0.5	1	2
0	0.200	0.500	1.000	2.000
1	0.164	0.303	0.368	0.271
2	0.134	0.184	0.135	0.007
3	0.110	0.112	0.050	0.005
4	0.090	0.067	0.018	0.000
5	0.073	0.041	0.007	0.000
6	0.060	0.025	0.002	0.000

It can be proved that :

1. the exponential density function decreases in the range 0 to ∞, the maximum ordinate of the curve occurring at $x = 0$ (the value being λ itself),

2. larger the value of λ, steeper is the decline in the ordinate, even for small values of x,

3. smaller the value of λ, flatter does the curve become and lies closer to X-axis.

The exponential distribution arises in practice when the random variable studied is servic time—the time taken to complete service at a filling station, grocery or automobile repair shop. It is also used in reliability theory to model the life times of components subject to wear, *e.g.*, batteries transistors, tubes, bulbs, etc. It has also been used to model the distribution of length of time between successive random events—the time between arrival of two customers at a service station or the time between breakdowns of a machine.

The mean and variance of the exponential distribution are given by

$$\text{Mean} = \frac{1}{\lambda}; \quad \text{Variance} = \frac{1}{\lambda^2}$$

MISCELLANEOUS ILLUSTRATIONS

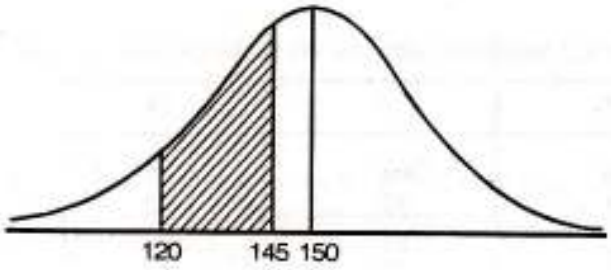
Illustration 21. The average daily sales of 500 branch offices was Rs. 150 thousand and the standard deviation Rs. 15 thousand. Assuming the distribution to be normal, indicate how many branches have sales between :

(*i*) Rs. 120 thousand and Rs. 145 thousand.

(*ii*) Rs. 140 thousand and Rs. 165 thousand.

Solution. (*i*) Standard normal variate corresponding to 120 is

$$z = \frac{X - \mu}{\sigma} = \frac{120 - 150}{15} = -2$$



and corresponding to 145, the standard normal variate is

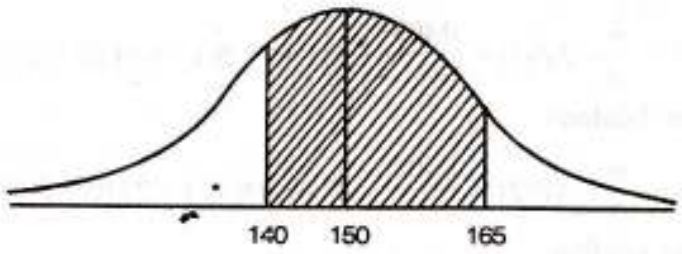
$$z = \frac{145 - 150}{15} = \frac{-5}{15} = -0.33.$$

From the table, we find the areas corresponding to the values of z arè 0.4772 and 0.1293.
Therefore, the desired area between Rs. 120 and Rs. 145 = 0.4772 – 0.1293 = 0.3479.
Hence, the expected number of branches having sales between Rs. 120 thousand and Rs. 145 thousand are $0.3479 \times 500 = 173.95 = 174$ approx.
(*ii*) Standard normal variate corresponding to 140 is

$$z = \frac{140 - 150}{15} = \frac{-10}{15} = -0.67.$$

and corresponding to 165, the standard normal variate is

$$z = \frac{165 - 150}{15} = 1$$



From the table, the areas corresponding to the z values are 0.2486 and 0.3413.
Therefore, the area is 0.2486 + 0.3413 = 0.5899
Hence, the expected number of branches having sales between Rs. 140 thousand and Rs. 165 thousand are $0.5899 \times 500 = 294.95$ or 295 approx.

Illustration 22. Eight coins are thrown simultaneously. Using binomial distribution, show that the probability of obtaining at least 6 heads is 0.1445. [*MBA, DU, 2003*]

Solution. If eight coins are thrown simultaneously, the probability of getting at least six heads will be given by the separate probabilities of getting 6 heads, 7 heads and 8 heads,

$$= P[x \geq 6] = P[x = 6] + P[x = 7] + P[x = 8]$$

$$n = 8; x = 6, 7, 8; \; p = \frac{1}{2} = q$$

$$= {}^8C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^2 + {}^8C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^1 + {}^8C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^0$$

$$= 28 \left(\frac{1}{2}\right)^8 + 8 \left(\frac{1}{2}\right)^8 + \left(\frac{1}{2}\right)^8$$

$$= \left(\frac{1}{2}\right)^8 [28 + 8 + 1] = \frac{37}{256} = 0.1445$$

Illustration 23. Five hundred television sets are inspected as they come off the production line and the number of defects per set is recorded below :

No. of defects (X):	0	1	2	3	4
No. of sets :	368	72	52	7	1

Estimate the average number of defects per set and expected frequencies of 0, 1, 2, 3 and 4 defects assuming Poisson distribution.

Solution. For finding the average number of defects, we make the following table :

X	f	fX
0	368	0
1	72	72
2	52	104
3	7	21
4	1	4
	$N = 500$	$\Sigma fX = 201$

Therefore, $$\overline{X} = \frac{\Sigma fX}{N} = \frac{201}{500} = 0.402$$

Thus, average number of defects per set is $m = 0.402$

The expected frequency of getting 0 defect

$$= NP(0) = 500 \times e^{-0.402}$$

But $e^{-0.402} = 0.6689$ [From the table]

Therefore, $NP(0) = 500 \times 0.6689 = 334.45$

Expected frequency of getting one defect

$$NP(1) = m.NP(0) = 0.402 \times 334.45 = 134.45$$

Expected frequency of getting 2 defects

$$NP(2) = \frac{m}{2}.\ NP(1) = \frac{0.402}{2} \times 134.45 = 201 \times 134.45 = 27.02$$

Expected frequency of getting 3 defects

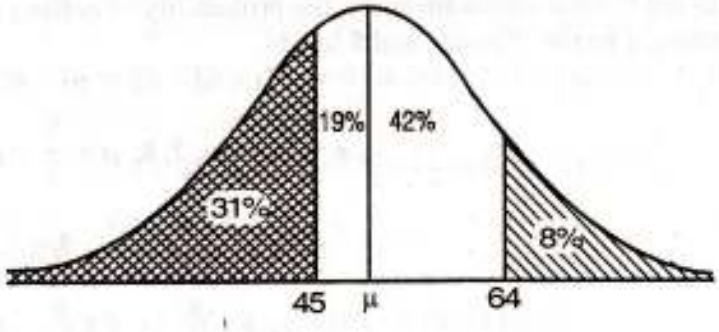
$$NP(3) = \frac{m}{3}.\ NP(2) = \frac{0.402}{3} \times 27.02 = 0.134 \times 27.02 = 3.62$$

Expected frequency of getting 4 defects

$$NP(4) = \frac{m}{4}.\ NP(3) = \frac{0.402}{4} \times 3.62 = 0.1005 \times 3.62 = 0.364.$$

Illustration 24. In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution. *(MBA, Delhi Univ., 1999)*

Solution. Let mean be μ and standard deviation σ, 31% of the items are under 45. They are lying to the left of the ordinate at $X = 45$ is 0.31, and therefore, are lying to the right of the ordinate up to the mean is $(0.5 - 0.31) = 0.19$. The value of z corresponding to this area is 0.5.



Hence $$z = \frac{45 - \mu}{\sigma} = -0.5 \quad ...(i)$$

8% of the items are above 64. Therefore, area to the right of the ordinate at 64 is 0.08. Area to the left of the ordinate at $X = 64$ up to mean ordinate is $(0.5 - 0.08) = 0.42$ and the value of z corresponding to this area is 1.4.

Hence
$$z = \frac{64-\mu}{\sigma} = 1.4 \qquad ...(ii)$$

From equations (*i*) and (*ii*)

$$-\mu + 0.5\sigma = -45$$
$$-\mu - 1.4\sigma = -64$$
$$1.9\sigma = 19 \quad \text{or } \sigma = 10$$
$$\mu - 0.5 \times 10 = 45 \quad \text{or } \mu = 50$$

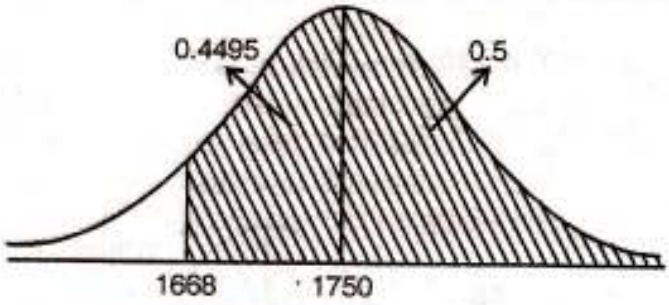
The mean of the distribution is 50 and standard deviation 10.

Illustration 25. The income of a group of 10,000 persons was found to be normally distributed with mean Rs. 1750 p.m. and standard deviation Rs. 50. Show that of this group 95% had income exceeding Rs. 1668 and only 5% had income exceeding Rs. 1832. What was the lowest income among the richest 100 ?

Solution. Standard normal variate is

$$z = \frac{X-\mu}{\sigma}$$

Here
$$X = 1668,\ \mu = 1750,\ \sigma = 50$$



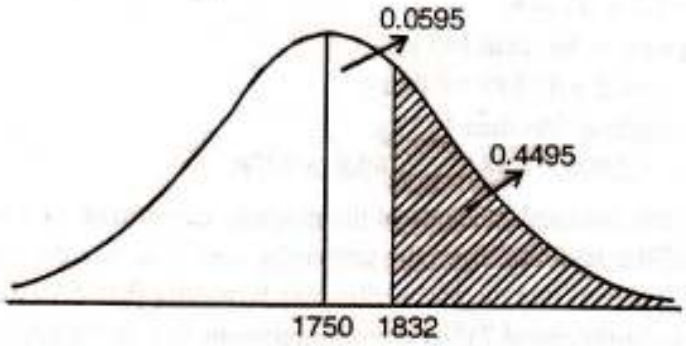
$$z = \frac{1668-1750}{50} = \frac{-82}{50} = -1.64$$

Area to the right of the ordinate at – 1.64 is (0.4495 + 0.5000) = 0.9495.

∴ The expected number of persons getting above Rs. 1668

$$= 10{,}000 \times 0.9495 = 9495$$

This is about 95% of the total, *i.e.*, 10,000.



The standard normal variate corresponding to 1832 is

$$z = \frac{1832-1750}{50} = \frac{82}{50} = 1.64$$

Area to the right of ordinate at 1.64 is

$$0.5000 - 0.4495 = 0.0505$$

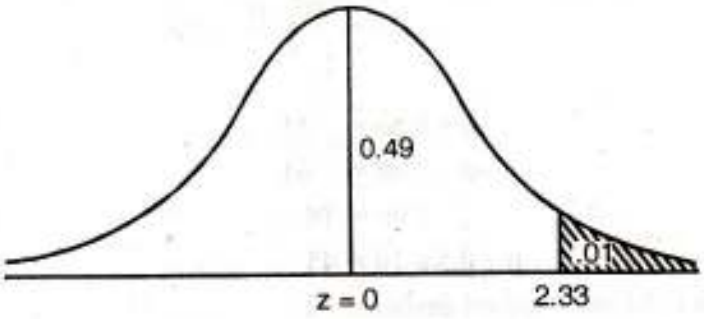
The number of persons getting above Rs. 1832 is

$$10{,}000 \times 0.0505 = 505$$

This is 5% approx. of the total, *i.e.*, 10,000.

Probability of getting richest 100

$$= \frac{100}{10{,}000} = 0.01$$



Standard normal variate having 0.01 area to its right = 2.33

$$2.33 = \frac{X - 1750}{50}$$

$$X = 2.33 \times 50 + 1750 = \text{Rs. } 1866 \text{ approx.}$$

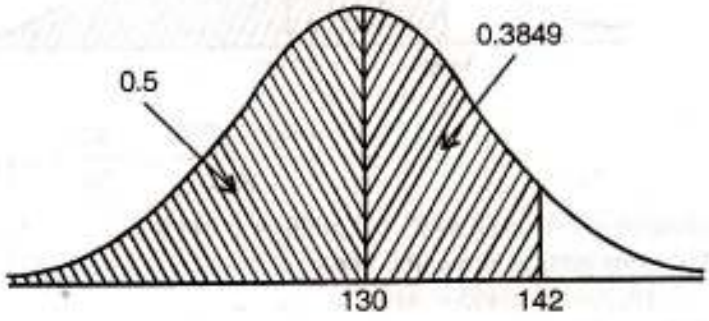
Hence, the lowest income among the richest 100 is Rs. 1866.

Illustration 26. A workshop produces 2000 units per day. The average weight of units is 130 kg with a standard deviation of 10 kg. Assuming normal distribution, how many units are expected to weigh less than 142 kg?

Solution. We are given :

$$\mu = 130,\ \sigma = 10,\ N = 2{,}000,\ X = 142$$

$$z = \frac{X - \mu}{\sigma} = \frac{142 - 130}{10} = 1.2$$



Area between $z = 0$ and $z = 1.2$ is 0.3849.

Probability of units having weight less than 142 kg

$$= 0.5 + 0.3849 = 0.8849$$

Expected number of units weighing less than 142 kg

$$= 2{,}000 \times 0.8849 = 1769.8 \text{ or } 1770.$$

Illustration 27. There are 600 business students in the graduate department of a university, and the probability for a student to need a copy of a particular textbook 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 probability law.)

Solution. Let X represent the number of copies of a textbook required on any day.

$$\text{Mean} = \mu = np = 600 \times 0.05 = 30$$

$$\text{s.d.} = \sigma = \sqrt{npq} = \sqrt{600 \times 0.05 \times 0.95} = \sqrt{28.5} = 5.3$$

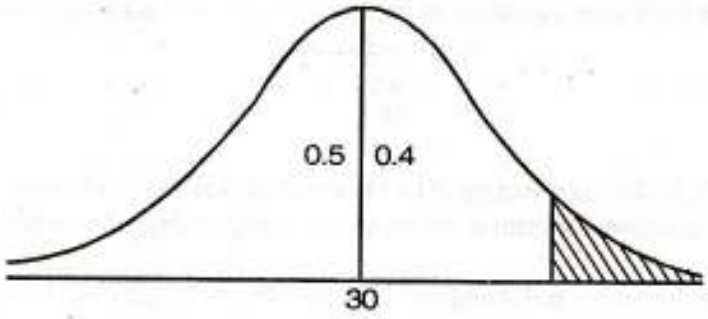
Therefore, area between mean and ordinate at X is greater than $(0.9 - 0.5)$, *i.e.*, 0.4, corresponding to this area, standard normal variate $z = 1.28$, and we get

$$\frac{X - \mu}{\sigma} > z$$

$$\frac{X - 30}{5.3} > 1.28$$

or

$$X - 30 > 6.784 \text{ or } X > 36.784$$



Hence, 37 copies of the textbook are required on any day.

Illustration 28. 1,000 tube lights with a mean life of 120 days are installed in a new factory, their length of life is normally distributed with standard deviation 20 days. (*i*) How many tube lights will expire in less than 90 days ? (*ii*) If it is decided to replace all the tube lights together, what interval should be allowed between replacements if not more than 10 per cent should fail before replacement ?

Solution. (*i*) $\mu = 120$, $\sigma = 20$, $X = 90$

Standard normal variate is

$$z = \frac{90 - 120}{20} = -1.5$$

Area of the curve at ($z = -1.5$) up to the mean ordinate = 0.4332

Area to the left of $-1.5 = 0.5 - 0.4332 = 0.0668$.

Number of tube lights expected to expire in less than 90 days

$$= 0.0668 \times 1000 = 66.8 = 67$$

(*ii*) The value of standard normal variate corresponding to an area 0.4(0.5 – 0.1) is 1.28.

$$\frac{X - 120}{20} = -1.28$$

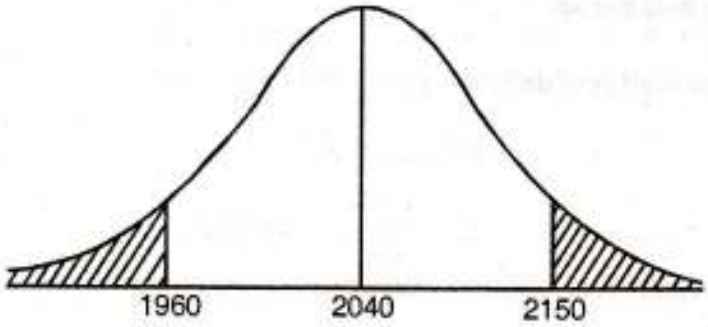
$X = 120 - 1.28 \times 20 = 120 - 25.6 = 94.4$ or 94.

Hence, 10 per cent of the tube lights will have to be replaced after 94 days.

Illustration 29. As a result of tests on 20,000 electric fans manufactured by a company, it was found that lifetime of the fan was normally distributed with an average life of 2,040 hours and standard deviation of 60 hours. On the basis of the information, estimate the number of fans that is expected to run for (*a*) more than 2,150 hours and (*b*) less than 1,960 hours.

Solution. (*a*) $X = 2150$, $\mu = 2040$, $\sigma = 60$.

$$z = \frac{X - \mu}{\sigma} = \frac{2150 - 2040}{60} = \frac{110}{60} = 1.833$$



Area to the right of ordinate at 1.833

$$= 0.5 - 0.4664 = 0.0336$$

The number of fans that is expected to run more than 2,150 hours

$$= 0.0336 \times 20{,}000 = 672.$$

(*b*) $X = 1960$, $\mu = 2040$, $\sigma = 60$.

$$z = \frac{1960 - 2040}{60} = -1.333$$

Area to the left of ordinate at $1.333 = 0.5 - 0.4082 = 0.0918$

∴ The number of fans that is expected to run for less than 1.960 hours

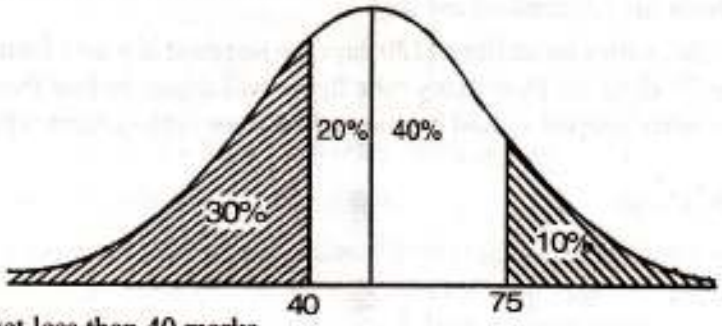
$$= 0.0918 \times 20{,}000 = 1836.$$

Illustration 30. The results of a particular examination are given below in a summary form :

Result	% of candidates
(*i*) Passed with distinction	10
(*ii*) Passed without distinction	60
(*iii*) Failed	30

It is known that a candidate fails in the examination if he obtains less than 40 marks (out of 100) while he must obtain at least 75 marks in order to pass with distinction. Determine the mean and standard deviation of the distribution of marks, assuming this to be normal.

Solution. We have to compute the mean and standard deviation from the given information. The following diagram will help in understanding the question and finding its solution :



We know that 30% students get less than 40 marks.

Therefore, from the table, z value corresponding to

$$0.2(20\% \text{ area}) = -0.524$$

Hence
$$\frac{40-\mu}{\sigma} = -0.524 \quad \ldots(i)$$

Also, 10% students get distinction marks, *i.e.*, 75 or more.

Therefore, from the table, z value corresponding to

$$0.4(40\% \text{ area}) = 1.28$$

Hence
$$\frac{75-\mu}{\sigma} = 1.28 \quad \ldots(ii)$$

Solving equations (*i*) and (*ii*), we get

$$\mu = 50.17 \text{ and } \sigma = 19.4$$

Hence, the mean of the distribution is 50.17 and standard deviation 19.4.

Illustration 31. A complex television component has 1,000 joints by machine which is known to produce on average, one defect in forty. The components are examined, and faulty soldering corrected by hand. If components requiring more than 35 corrections are discarded, what proportion of the components will be thrown away? (Apply normal distribution.)

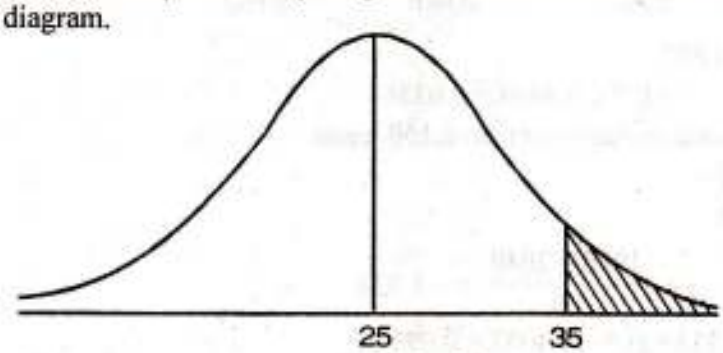
Solution. Since the probability of occurrence of the defect is very small, therefore, it is more appropriate to use normal distribution as a limiting case of Poisson distribution.

Here,
$$m = \text{average number of defects} = np = 1000 \times \frac{1}{40} = 25$$

$$\mu = \sqrt{m} = \sqrt{25} = 5$$

We know that
$$z = \frac{x-m}{\sqrt{m}} \sim N(0, 1).$$

Since it is required to find out the components requiring more than 35 corrections which have to be discarded is shown below by the shaded area in the diagram.



Therefore,
$$z = \frac{35-25}{5} = 2.$$

From the table, the corresponding value of $z = 2$ is 0.4772. The required shaded area will be $0.5 - 0.4772 = 0.0228$. Hence, the number of proportion of the components which will be thrown away is 2.28%.

Illustration 32. A baker has studied his record and notices that for the past 310 working days in the year, the demand for his product (bread) has varied as follows :

Demand ('000 units)	:	5	6	7	8	9	10
Number of days	:	20	60	80	120	20	10

What is the expected demand for his product ?

Solution. Let the demand ('000 units) be denoted by X. Then

$$E(X) = \Sigma X\, P(X)$$

$$= 5 \times \frac{20}{310} + 6 \times \frac{60}{310} + 7 \times \frac{80}{310} + 8 \times \frac{120}{310} + 9 \times \frac{20}{310} + 10 \times \frac{10}{310}$$

$$= \frac{1}{310}\,[100 + 360 + 560 + 960 + 180 + 100]$$

$$= \frac{1}{310}\,[2260] = \frac{226}{31} = 7.29032$$

Therefore, the expected demand $= 7.29032 \times 1000 = 7290.32$.

Illustration 33. In a town, 10 accidents took place in a span of 50 days. Assuming that the number of accidents per day follows Poisson distribution, find the probability that there will be three or more accidents in a day.

Solution. The average number of accidents per day is

$$m = \frac{10}{50} = 0.2$$

Prob. (3 or more accidents) = 1 – Prob. (2 or less accidents)

= 1– [Prob. (0 accident) + Prob. (1 accident) + Prob. (2 accidents)]

$$= 1 - [e^{-m} + me^{-m} + \frac{m^2}{2} e^{-m}]$$

$$= 1 - e^{-m}[1 + m + \frac{m^2}{2}]$$

$$= 1 - e^{-0.2}[1 + 0.2 + \frac{0.04}{2}]$$

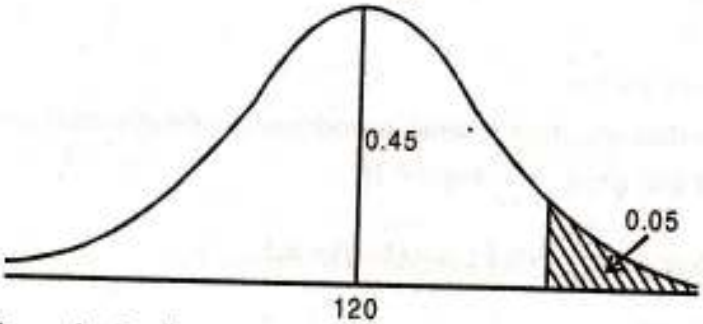
$$= 1 - 0.8187\ [1.22] = 1 - 0.9988 = 0.0012.$$

Illustration 34. A wholesale distributor of fertilizer products finds that the annual demand for one type of fertilizer is normally distributed with a mean of 120 tonnes and standard deviation of 16 tonnes. If he orders only once a year, what quantity should be ordered to ensure that there is only a 5 per cent chance of running short ? *(MBA, Delhi Univ., 2005, 2007)*

Solution. Let the annual demand (in tonnes) be denoted by the random variable X.

Therefore,

$$z = \frac{X - 120}{16}$$



The desired area of 0.05 is shown in the figure. Since the area between the mean and the given value of X is 0.45, therefore, from the table we get this area of 0.45 corresponding to $z = 1.64$.

Substituting this value of $z = 1.64$ in standardized normal variate, we get

$$1.64 = \frac{X - 120}{16}$$

or

$$X = 120 + (1.64)\,(16) = 146.24.$$

If it is necessary to order in whole units, then the wholesale distributor should order 147 tonnes.

Illustration 35. The probability that a bulb will fail before 100 hours is 0.2. Bulbs fail independently. If 15 bulbs are tested for life lengths, what is the probability that the number of failures before 100 hours does not exceed 3?

Solution. The given problem can be assumed to follow binomial distributions.

Here, $n = 15 ; p = 0.2 ; q = 0.8.$

The required probability is given by :

$$P[X \le 3] = P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3]$$
$$= {}^{15}C_0(0.2)^0(0.8)^{15} + {}^{15}C_1(0.2)^1(0.8)^{14} + {}^{15}C_2(0.2)^2(0.8)^{13} + {}^{15}C_3(0.2)^3(0.8)^{12}$$
$$= (0.8)^{15} + 15(0.2)^1(0.8)^{14} + 105(0.2)^2(0.8)^{13} + 455(0.2)^3(0.8)^{12}$$
$$= (0.8)^{12}[(0.8)^3 + 15(0.2)(0.8)^2 + 105(0.2)^2(0.8) + 455(0.2)^3]$$
$$= 0.0687[0.512 + 1.92 + 3.36 + 3.64]$$
$$= 0.0687(9.432) = 0.648.$$

Illustration 36. A manufacturer who produces medicine bottles, finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of bottles.

Using Poisson distribution, find how many boxes will contain :

(*i*) no defectives,

(*ii*) at least two defectives. *(MBA, Delhi Univ., 200[illegible])*

(Given $e^{-0.5} = 0.6065$)

Solution. We are given : $p = 0.001$, $n = 500$, $m = np = 500 \times 0.001 = 0.5$

(*i*) $P[X = 0] = e^{-m} = e^{-0.5} = 0.6065$

Therefore, the required number of boxes

$$= 0.6065 \times 100 = 60.65 \text{ or } 61$$

(*ii*) $P(X > 2) = 1 - [P[X = 0] + P[X = 1]]$

$$= 1 - [e^{-m} + me^{-m}]$$
$$= 1 - [0.6065 + 0.5(0.6065)]$$
$$= 1 - 0.6065 + 0.30325$$
$$= 1 - 0.90975 = 0.09025.$$

Therefore, the required number of boxes

$$= 100 \times 0.09025 = 9.025 \text{ or } 9.$$

Illustration 37. (*a*) Bring out the fallacy, if any, in the following statement. "The mean of a binomial distribution is 10 and its standard deviation is 6".

(*b*) The mean of a binomial distribution is 20 and the standard deviation is 4. Calculate n, p and q.

Solution. (*a*) The mean of a binomial distribution is np and standard deviation $\sqrt{npq}$.

$$np = 10 \text{ and } \sqrt{npq} = 6$$

Squaring, $npq = 36$ —(*i*)

Putting the value of np in (*i*)

$$10q = 36 \text{ or } q = 3.6$$

Since the value of q is greater than one, there is some inconsistency in the statement given.

(*b*) Given $np = 20$ and $\sqrt{npq} = 4$ or $npq = 16$ —(*ii*)

$$20q = 16 \text{ or } q = \frac{16}{20} = 0.8 ; \; p = (1 - q) = 0.2$$

Putting the value of p and q in (*ii*)

$$n(0.8)(0.2) = 16 \text{ or } n = 100.$$

Illustration 38. One-fifth per cent of the blades produced by a blade manufacturing company turn out to be defective. The blades are supplied in packets of 10. Use poisson distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades respectively in a consignment of 1,00,000 packets. *(MBA, DU, 2003; 2006)*

Solution. Given $p = \frac{1}{500}$, $n = 10$

$$np = 10 \times \frac{1}{500} = 0.02 = m$$

(*i*) Probability of no defective blade :

$$f(0) = Pr\,[X = 0] = e^{-m} = e^{-0.02} = 0.9802 \qquad [\text{Table value of } e^{-0.02} = 0.9802]$$

Therefore, number of packets containing no defective blade is given as :

$$Nf(0) = 1{,}00{,}000 \times 0.9802 = 98020.$$

(*ii*) Probability of one defective blade :

$$f(1) = Pr\,[X = 1] = me^{-m} = 0.02 \times 0.9802 = 0.019604$$

Therefore, approximate number of packets containing one defective blade is as :

$$Nf(1) = 1{,}00{,}000 \times 0.019604 = 1960.4 \simeq 1960.$$

(*iii*) Probability of two defective blades :

$$f(2) = Pr\,[X = 2] = \frac{m^2 e^{-m}}{2!} = \frac{(0.02)^2\,(0.02)}{2}$$

$$= (0.0004)\,(0.4901) = 0.00019604$$

Therefore, approximate number of packets containing two defective blades is as :

$$Nf(2) = 1{,}00{,}000 \times 0.00019604 = 19.6 \simeq 20$$

Illustration 39. A market researcher at a major automobile company classified households by car ownership. The relative frequencies of the households for each category of ownership are shown in the table :

Number of cars per household	*Relative frequencies*
0	0.10
1	0.30
2	0.40
3	0.12
4	0.06
5	0.02

(*a*) Establish the probability distribution for the random variable.

(*b*) Calculate the expected value of the random variable, and interpret the result.

(*c*) Compute the values of the variance and standard deviation of the probability distribution. *[MBA, DU, 2003]*

Solution. Let the random variable X denote number of cars per household. Therefore, the table shown below gives the required computation data.

No. of Cars/household X	*Rel. Freq.* $P(X)$	$XP(X)$	$X^2 P(X)$
0	0.10	0.00	0.00
1	0.30	0.30	0.30
2	0.40	0.80	1.60
3	0.12	0.36	1.08
4	0.06	0.24	0.96
5	0.02	0.10	0.50
		1.80	4.44

(*a*) The probability distribution for the random variable X is given in the first two columns of the above table.

(*b*) Expected value of the random variable X :

$E(X) = \Sigma X\ P(X) = 1.80$

This expected value is interpreted as that, on the average, there are 1.8 cars per household.

(*c*)
$$\text{Variance} = \sigma^2 = \text{Var}\,(X) = E(X^2) - [E(X)]^2$$
$$= \Sigma X^2\ P(X) - [1.80]^2 \qquad [E(X) = 1.80]$$
$$= 4.44 - 3.24 = 1.20$$

$$\text{Standard deviation} = \sigma = \sqrt{1.20} = 1.095.$$

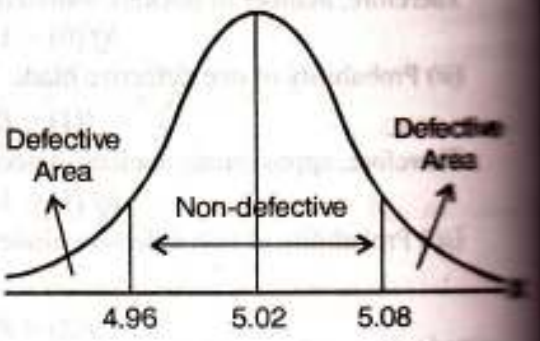
Illustration 40. The mean inside diameter of a sample of 500 washers produced by a machine is 5.02 mm, and the standard deviation is 0.05 mm. The purpose for which these washers are intended, allows a maximum tolerance in the diameter of 4.96 to 5.08 mm., otherwise the washers are considered defective. Determine the percentage of defective washers produced by the machine assuming the diameters are normally distributed. *(MBA, DU, 2003)*

Solution. Given : $\mu = 5.02, \sigma = 0.05$

Using the Area under the normal curve, the standard normal variate

$$z_1 = \frac{X - \mu}{\sigma}, \text{ therefore } z_1 = \frac{5.08 - 5.0}{0.05} = 1.2$$

and $$z_2 = \frac{4.96 - 5.08}{0.05} = -1.2$$



For value of $z = 1.2$, the correspond area from the table is 0.3849. Therefore, the non-defective washer are 0.3849 + 0.3849 = 0.7698. Hence, the probability of defective washers produced by the machine is 1 – 0.7698 = 0.2302.

Therefore, the percentage of defective washers is 23.02.

Illustration 41. The demand for a product of a company varies greatly from month to month. The probability distribution in the following table, based on the past 2 years of data, shows the Company's monthly demand :

Unit Demand	*Probability*
300	0.20
400	0.30
500	0.35
600	0.15

(*i*) If the company bases monthly orders on the expected value of the monthly demand, what should monthly order quantity be for this product ?

(*ii*) Assume that each unit demanded generates Rs. 70 in revenue and that each unit ordered costs Rs. 50. How much will the company gain or loss in a month if it places an order based on your answer to part (*i*) and the actual demand for the item is 300 units ?

Solution : (*i*)

Unit Demand X	*Probability* $P(X)$	$XP(X)$
300	0.20	60
400	0.30	120
500	0.35	175
600	0.15	90
		$\Sigma XP(X) = 445$

Expected Value : $E(X) = \Sigma XP(X) = 445$

Therefore, monthly order quantity for the product is 445 units.

(*ii*) Profit = Revenue – Cost = 70 – 50 = Rs. 20

Profit for 445 units = 445 × 20 = Rs. 7250.

Profit for 300 units = 300 × 20 = Rs. 6000

Company Loss = Rs. 7250 – Rs. 6000 = Rs. 1250

Illustration 42. A new automated production process has had an average of 1.5 breakdowns per day. Because of the cost associated with a breakdown, management is concerned about the possibility of having three or more breakdowns, during a day. Assume that breakdowns occur randomly, that the probability of a breakdown is the same for any two time intervals of equal length, and that breakdowns in one period are independent of breakdowns in other periods. What is the probability of having three or more breakdowns during a day ? *(MBA, Delhi Univ., 2003)*

Solution : Given : $m = 1.5, e^{-m} = e^{-1.5} = 0.2231$

$$f(x \geq 3) = 1 - [Pr(x = 0) + Pr(x = 1) + Pr(x = 2)]$$

$$= 1 - [e^{-m} + me^{-m} + \frac{m^2}{2} e^{-m}]$$

$$= 1 - e^{-m}[1 + m + \frac{m^2}{2}]$$

$$= 1 - 0.2231\,[1 + 1.5 + \frac{(1.5)^2}{2}]$$

$$= 1 - 0.2231\,[3.625] = 1 - 0.8088 = 0.1912.$$

Assume that the test scores from a college admissions test are normally distributed with a mean of 450 and ... of 100.

... percentage of the people taking the test score are between 400 and 500 ?

... someone received a score of 630. What percentage of the people taking the test score better ? What percentage ... worse ?

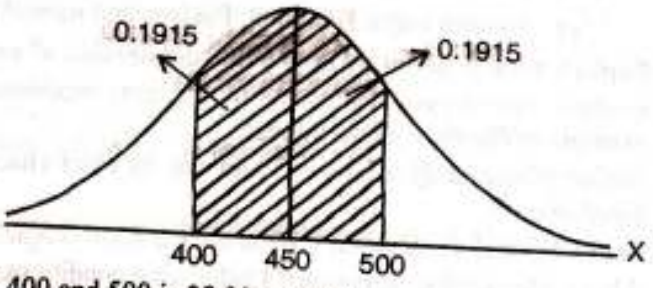
... particular university will not admit anyone scoring below 480, what percentage of the persons taking the test would be acceptable to the university ?

Solution : Let X denote the test scores.

$$X \sim N\ (450,\ 100)$$

(*i*) $$z_1 = \frac{X-\mu}{\sigma} = \frac{500-450}{100} = 0.5$$

$$z_2 = \frac{400-450}{100} = -0.5$$



Corresponding to $z = 0.5$, the area is 0.1915.

The required probability = $0.1915 \times 2 = 0.3830$

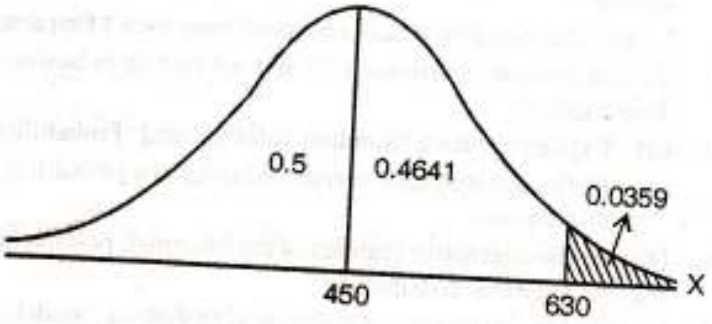
Hence, percentage of the people taking the test score between 400 and 500 is 38.30 per cent.

(*ii*) $$z = \frac{630-450}{100} = 1.8$$

Corresponding to $z = 1.8$, the area is 0.4641.

The required prob = $0.5 - 0.4641 = 0.0359$ for test score better and required probability $= 1 - 0.0359 = 0.9641$.

Hence, only 3.59 per cent are better and 96.41 per cent are worse.

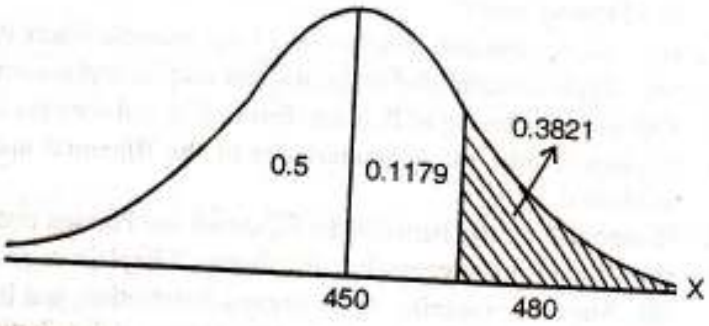


(*iii*) $$z = \frac{480-450}{100} = 0.30$$

For $z = 0.30$, the area is 0.1179

required Prob. = $0.5 - 0.1179 = 0.3821$

Therefore, the per centage of the persons taking the test ... acceptable is 38.21.



PROBLEMS

I-A : Answer the following questions, each question carries **one** mark:

(*i*) Define binomial distribution ?

(*ii*) What is Bernoulli trial ? *(MBA, Hyderabad Univ., 2006)*

(*iii*) What is normal distribution ?

(*iv*) State and explain the constants of binomial distribution ? *(MBA, Madurai-Kamaraj Univ., 2003; M.A. Eco., M.K. Univ., 2003)*

(*v*) Write the parameters of binomial distribution.

(*vi*) Write the distribution in which mean and variance will be the same.

(*vii*) Mean value of binomial distribution is .....

(*viii*) Define normal distribution and list its properties.

(*ix*) What is normal probability distribution ? *(MBA , UP Tech. Univ., 2006)*

(*x*) Mean value of Poisson distribution is .....

(*xi*) What is Poisson distribution ?

(*xii*) Explain any two properties of normal distribution.

1-B : Answer the following questions, each question carries **four** marks:

(*i*) Distinguish between

(*i*) The mean of a binomial distribution is 20 and standard deviation is 4. Find out n, p, and q. *(M. Com., M.K. Univ., 2002)*

(*ii*) Define Poisson distribution and state its uses.

(*iii*) Define Binomial distribution. The parameters of a binomial distribution are $n = 10$ and $p = 0.2$. Find the mean and variance of the distribution. *(MBA, Madras Univ., 2003)*

(*iv*) Give at least five important properties of normal distribution.

(*v*) In what ways, Binomial, Poisson and normal distribution are related.

2. Explain, what is meant by Probability distribution of a discrete random variable.
3. Explain, what do you understand by the term 'mathematical expectation'. How is it useful for a businessman ? Give an example to illustrate its usefulness. *(MBA, Delhi Univ., 2004)*
4. Define Binomial distribution. Point out its chief characteristics and uses. Under what conditions, it tends to Poisson distribution ? *(MBA, Osmania Univ., 2002)*
5. Define Normal distribution ? What are the main characteristics of Normal Distribution?
6. What is Binomial distribution ? Under what conditions will it tend to Normal distribution?
7. What are the chief properties of Normal distribution ? Describe briefly the importance of Normal distribution in statistical analysis.
8. Under what conditions, can observed (empirical) frequency distributions be approximated to binomial distribution ?
9. What is Poisson distribution ? Point out its role in business decision-making. Under what conditions will it tend to normal distribution ? *(MBA, Kumaun Univ., 2006)*
10. (*a*) Explain the term 'Random variable' and 'Probability distribution of a random variable'.
    (*b*) Define the Binomial variate and obtain its probability distribution function. Find the mean and variance of the binomial distribution.
11. Discuss the distinctive features of the binomial, poisson and normal distribution. When does a binomial distribution tend to become a normal distribution ?
12. Under what conditions is the binomial probability model appropriate ? How does it approach the Poisson probability model as a limiting case ?
13. (*a*) What is Poisson distribution ? Give example where it can be applied.
    (*b*) Explain binomial distribution and give its application in business management.
14. Explain the meaning of Poisson distribution and state the conditions under which this distribution is used.
15. Explain briefly, the characteristics of the Binomial and Poisson distributions. How are their means and variances calculated ?
16. Distinguish clearly between the Binomial and Poisson distribution. *(MBA, UP Tech. Univ., 2007)*
17. (*a*) What is hypergeometric distribution ? Explain its properties.
    (*b*) State the properties of the normal distribution, and Binomial distribution. *(MBA, Hyderabad Univ., 2005)*
    (*c*) Describe briefly, the importance of normal distribution in business decision-making. What are its chief properties?
18. (*a*) Briefly describe the characteristics of the normal probability distribution. Why does it occupy such a prominent place in statistics ?
    (*b*) When can Poisson distribution be a reasonable approximation of the binomial?
    (*c*) Fifty per cent of all automobile accidents lead to property damage of Rs. 100. Forty per cent lead to damage of Rs. 500. Ten per cent lead to total loss, a damage of Rs. 1,800. If a car has a 5 per cent chance of being in an accident in a year, what is the expected value of the property damage due to that possible accident ?
19. Suppose that in a lottery 1,000 tickets are sold at Rs. 10 each, and three prizes are to be awarded. The first prize is a television set worth Rs. 12,000, second prize is a short wave radio worth Rs. 1500; and the third prize is a cycle worth Rs. 1300. If you plan to buy one ticket, what is your expected gain or loss from the venture ?
20. Two investment opportunities are open to prospective investor. If opportunity A turns out to be successful, a profit of Rs. 6 lakh will result and the probability of A's success is estimated as 0.75, if A turns out to be a failure there will be a loss of Rs. 1 lakh. If opportunity B succeeds a profit of Rs. 25 lakh will materialize but, if it fails, there will be a loss of Rs. 7 lakh and the probability for B to fail is 0.55. Which investment opportunity should the investor take if the dicision criterion is to maximize profits?
    [Opportunity B]
21. If it rains, a raincoat dealer can earn Rs. 5000 per day. If it is fair, he can lose Rs. 1000 per day. What is his expectation if the probability of rain is 0.4 ?
    [Rs. 1400]

22. A firm plans to bid Rs. 3000 per tonne for a contract to supply 1,000 tonnes of a metal. It has two competitors A and B and it assumes that the probability that A will bid less than Rs. 3000 per tonne is 0.3 and that B will bid less than Rs. 3000 per tonne is 0.7. If the lowest bidder gets all the business and the firms bid independently, what is the expected value of the contract to the firm ?
[(*i*) 0.18; (*ii*) 0.32]

23. If the probability that an individual suffers from reaction of a given medicine is 0.001, determine the probability that out of 2,000 individuals (*i*) exactly 3 individuals (*ii*) more than 2 individuals will suffer from reaction.
[(*a*) 0.1353, (*b*) 0.2706, (*c*) 0.2706, (*d*) 0.1804]

24. If 2% of the electric bulbs manufactured by a company are defective, find the probability that in a sample of 100 bulbs (*a*) 0, (*b*) 1, (*c*) 2 and (*d*) 3 bulbs will be defective.
[0.3601]

25. A certain type of plastic bag in the past has burst under a pressure of 10 pounds 30% of the time. If a prospective buyer tests 5 bags chosen at random, what is the probability that exactly one will burst ?
[0.302]

26. The probability that A will make a profit on any business deal is 0.8, what is the probability that he will make a profit exactly eight times in ten successive deals ?

27. The distribution of typing mistakes committed by a typist is given below. Assuming a Poisson model, find out the expected frequencies.

Mistakes per page :	0	1	2	3	4	5
No. of pages :	142	156	69	27	5	1

[147.12, 147.12, 73.56, 24.52, 6.13, 1.22] *(MBA, Sukhadia Univ., 2004; Hyderabad Univ., 2006)*

28. In a normal distribution, 7% of the observations are under 35 and 89% are under 63. What are the mean and the standard deviation of distribution ?
[$\mu = 50.28$, $\sigma = 10.25$]

29. If the average number of rejects in the manufacturing process of a certain article is 4 per cent, what are the probabilities of 0, 1, 2, 3, 4 rejects in a sample of 40 articles ?
[0.6703, 0.2681, 0.0536, 0.00715, 0.000715]

30. A Municipal Corporation had installed 5,000 bulbs in the streets of the city. If these bulbs have an average life of 800 burning hours, with a standard deviation of 200 hours, find :
(*i*) What number of bulbs might be expected to fail in the first 600 burning hours ? (*ii*) The number of bulbs expected to fail between 700 and 900 burning hours, and (iii) the number of bulbs expected to fail after 900 burning hours.
[(*i*) 794, (*ii*) 1915, (*iii*) 1542.5]

31. The weekly wages of 1,000 workers are normally distributed with a mean of Rs. 1700 and a standard deviation of Rs. 150. Estimate the lowest weekly wages of the 100 highest paid workers.

32. Find the probability that at most 5 defective bolts will be found in a box of 200 bolts, if it is known that 2 per cent of such bolts are expected to be defective. (You may take the distribution to be Poisson.)
[0.784]

33. If 20% of the bolts produced by a machine are defective, determine the probability that out of 4 bolts (*i*) 0, (*ii*) 1 and (*iii*) at the most 2 bolts will be defective.
[0.4096, 0.4096, 0.9728]

34. The probability that any customer who enters the store will purchase Colgate toothpaste is 0.3. If 1,000 customers enter the store, what is the minimum number of Colgate toothpastes the store must have on hand, if the probability that it will be out of stock is to be at most 1% ?
[334]

35. Daily demand for a product is approximately normally distributed with mean sales of 12 units per day and standard deviation of 4 units. How many units must be on hand in the morning to assure no more than one chance in 5 of running out of stock during the day ? *(MBA, DU, 2003, 2006)*
[16]

36. The probability that India wins a cricket Test match against England is given to be 1/3. If India and England play three Test matches, use binomial distribution to find the probability that :
(*i*) India will lose all three Test matches ? (*ii*) India will win at least one Test match ?
[(*i*) 0.2963, (*ii*) 0.7037]

37. An individual is offered an opportunity to bet Rs. 500 on the outcome of a roll of a pair of dice. If the dice turn up so that the sum of the faces total 7 or 11, the individual wins Rs. 1500. For any other outcome the bet is lost. What is the expected value of the game for the individual ?

38. The fuel consumption of a fleet of 150 trucks is normally distributed with a mean of 15 km per litre and a standard deviation of 1.5 km per litre. Use normal distribution to find the expected number of trucks that average :
(*a*) 13 but less than 14 km per litre, (*b*) 14.5 but less than 15.5 km per litre.
[(*a*) 24, (*b*) 39]

39. Fit a Poisson distribution to the data given below :

X :	0	1	2	3	4
f :	123	59	14	3	1

40. The heights of students in a class are normally distributed with a mean of 62 inches and a standard deviation of 4 inches. What proportion of the students in the class have a height greater than 68 inches ? What is the probability that a student selected at random will have a height between 58 inches and 66 inches ?

41. One hundred car radio sets are inspected as they come off the production line and number of defects per set is recorded below :

No. of defects :	0	1	2	3	4
No. of sets :	79	18	2	1	0

Fit a Poisson distribution to the above data.
[77.88, 19.47, 2.43, 0.2028, 0.1267] ***(MBA, Delhi Univ., 2006)***

42. A machine is supposed to drill holes with a diameter of 1 inch. In fact, the diameters are normally distributed with a mean of 1.0 inches and a standard deviation of 0.02 inch. If there is a tolerance of 0.02 inch, the holes should be between 0.99 and 1.02 inches. What percentage of the holes drilled are within tolerance limits ?

43. A hotel maintains two delux rooms. The demand for these rooms in any day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of day on which neither of the rooms would be used ; and the proportion of days on which no demand would be refused.
[0.2231, 0.1913]

44. The following frequency table gives the distribution of 1,000 persons according to their income :

Monthly income (Rs.)	*No. of persons*	*Monthly income (Rs.)*	*No. of persons*
Below 5000	16	20000–25000	166
5000–10000	85	25000–30000	100
10000–15000	207	30000–35000	69
15000–20000	346	above 35000	11

Fit a normal distribution to the above frequency table. Also, determine the percentage of persons with an income between 17,500 and 27,500 rupees.

45. A car rental firm has two cars which it rents out day by day. The number of demand for a car on each day is distributed as Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused.

46. 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 limits between which the probability of a throw of 3 or 4 is expected to lie.

47. An automatic detergent packing machine produces packages whose weights are normally distributed with a mean of 8.00 gm and a standard deviation of 0.010 gm.
(*a*) What proportion of packages are between 7.98 gm and 8.000 gm ?
(*b*) What proportion are between 8.005 and 8.0151 gm ?
(*c*) What proportion are between 7.995 and 8.010 gm ?
(*d*) What proportion are above 8.017 gm ?
[(*a*) 47.72 (*b*) 24.30 (*c*) 53.28 (*d*) 4.46.]

48. A manufacturer of electric fuses packs fuses in boxes of 10 each and 2,000 such boxes were sold. The previous experience shows that 5 per cent of the fuses are defective. Using Poisson distribution, find how many boxes will contain (*i*) no defective (*ii*) more than one defective.

49. In a certain factory turning out razor blades, there is a small chance of 1/500 for any blade to be defective. The blades are supplied in packets of 10. Use Poisson distribution to calculate the number of packets containing (*i*) no defective (*ii*) one defective and (*iii*) two defective blades, respectively in a consignment of 10,000 packets.

50. If, on an average 8 ships out of 10 arrive safely at port, find the mean and standard deviation of the number of ships arriving safely out of a total of 1,600 ships.
[$\mu = 1280$, $\sigma = 16$]

51. Assuming that sex ratio of male children is $\frac{1}{2}$, find the probability that in a family of 5 children, (*i*) all children will be of the same sex, and (*ii*) three of them will be boys and two girls.
[(*i*) 1/16 (*ii*) 5/16]

52. A company has 6 telephones which 10 executives use intermittently. Assume that at any given time each executive has the same probability 'p' of requiring to use a telephone. If the executives' requirements of telephones are independent, the probability that exactly k executives require a phone is $b\ (k, n, p)$. If on an average, an executive uses the telephone for 10 minutes per hour ($p = 1/6$), find the probability that 7 or more executives need a telephone at the same time.
[0.000267]

53. A machine produces bolts which are 10% defective. Find the probability that in random sample of 400 bolts produced by this machine, the number of defectives found
(*i*) will be at most 30;
(*ii*) will be between 30 and 50;
(*iii*) will exceed 55.
[(*i*) 22.8, (*ii*) 354.32, (*iii*) 3.56]

54. The mean and standard deviation for the life times of a population of light bulbs are 1200 and 150 hours respectively. Assuming these lifetimes are normally distributed, what is the probability that a light bulb will last over 1500 hours ?
[0.0228]

55. An editor of a publishing company, calculates that it requires 11 months on an average to complete the publication process from manuscript to finished books with a standard deviation of 2.4 months. He believes that the distribution of publication times is well described by the normal distribution. Out of 190 books he will handle this year, how many will complete the process in less than a year ?
[126]

56. An analyst predicts that 2.5% of all small companies will file for bankruptcy in the coming year. For a random sample of 200 companies, estimate probability that
(*i*) at least three will file for bankruptcy next year ;
(*ii*) exactly three will file for bankruptcy ;
(*iii*) not more than five will file for bankruptcy.
[(*i*) 0.87 (*ii*) 0.145 (*iii*) 0.6151]

57. Past records show that the average number of accidental drownings at a beach resort is 3 per year for every 100,000 of tourists visiting the resort. If in a year 200,000 tourists visited this resort, find the probabilities that :
(*i*) there will be no drowning accident this year ;
(*ii*) there will be at least 6 accidents this year ;
(*iii*) there will be exactly 5 accidents this year ;
(*iv*) there will be more than 8 accidents this year.
[(*i*) 0.00279 (*ii*) 0.498 (*iii*) 0.1807 (*iv*) 0.0465]

58. Fit a binomial distribution to the following data :

X:	0	1	2	3	4
f:	28	62	28	12	46

[9.34, 40.48, 65.78, 47.51, 12.86]

59. The financial controller of Galaxy Airlines is having some problems with cash flows. Daily revenue fluctuate greatly and are difficult to predict whereas daily expenses remain fairly constant regardless of the daily number of passengers. If daily revenue has a normal distribution with a mean of Rs. 72,000 and 85 per cent of the values lie below Rs. 82,000, what is the standard deviation of the distribution. What is the value above which 5 per cent of the values in the distribution lie ?

60. The following table shows the number of customers returning the products in a marketing territory. The data is for 100 stores :

No. of returns :	0	1	2	3	4	5	6
No. of stores :	4	14	23	23	18	9	9

Fit a Poisson distribution.
[4.97, 14.91, 22.36, 22.36, 16.77, 10.06, 5.03]

61. Three fair coins are tossed 300 times. Find the frequencies of the distribution of heads and tails and tabulate the result. Also, calculate the mean and standard deviation of the distribution.
[2.25, 1.06]

62. How many workers have a salary above Rs. 2,675 in the distribution whose average salary is Rs. 2,400 and standard deviation is Rs. 100 and the number of workers in the factory is 15,000, if the salary of workers follows the normal law ?
[4368]

63. The Delhi Municipal Corporation installed 2,000 bulbs in the streets of Kailash Colony. If these bulbs have an average life of 1,000 burning hours, with a standard deviation of 200 hours, what number of bulbs might be expected to fail in the first 700 burning hours?
[134] *(MBA, HPU)*

64. In 24 trials of an event of small probability, the frequency f of the number of success X is given in the following table :

X :	0	1	2	3	4	5	6
f :	3	2	6	5	5	1	2

The mean number of successes is 2.75. Find the expected frequencies of the Poisson distribution with the same total frequency.

65. The appearing of 2 or 3 on a dice is connected as success. Five dice are thrown 729 times and the following results are observed :

Number of successes :	0	1	2	3	4	5
Frequency :	45	195	237	132	81	39

Fit the binomial distribution assuming the dice to be unbiased.
[96, 240, 240, 120, 30, 3]

66. Five hundred T.V. sets are inspected as they come off the production line and average number of defects per set is found to be 0.402. Find the expected number of T.V. sets having one or more defects.
[165.5]

67. The time required by bank cashier to deal with a customer has been observed to be normally distributed with mean 25 secs. and a standard deviation 10 secs. Find the probability that a customer arriving at random will have to wait :
(*i*) between 20 and 28 secs. ;
(*ii*) less than 23 secs.
To what value should the mean service time be altered so that only 1 customer in 100 has to wait longer than 50 secs. ?
(*i*) 0.3094 (*ii*) 0.4207, 26.7

68. The marks obtained by the students in an examination are known to be normally distributed. If 10% of the students got less than 40 marks while 15% got over 80, what are the mean and standard deviation of marks ? *(MBA, Delhi Univ., 2006)*
[$\mu = 62.0688$, $\sigma = 17.2413$]

69. In a certain examination, 10% of the students got less than 30 marks and 97% of the students got less than 62 marks. Assuming the distribution to be normal, find the mean and standard deviation of the marks.
[$n = 42.96$, $a = 10.13$]

70. For a binomial distribution the mean is 4 and variance 2. Find the probability of getting
(*i*) at least 2 successes.
(*ii*) at most two successes. *(MBA., DU, 2002, 2007)*
[(*i*) 0.9648 (*ii*) 0.1445]

71. The following table gives the numbers of days in a 50 day period in which automobile accidents occurred in a certain part of a city. Fit a Poisson distribution to the data.

No. of accidents :	0	1	2	3	4
No. of days :	19	18	8	4	1

(18.4, 18.4, 9.2, 3.1, 0.8]

72. In a book, the following frequency of mistakes per page was observed. Fit a Poisson distribution.

No. of mistakes per page :	0	1	2	3	4	5
No. of pages :	630	160	90	70	30	20

[463, 357, 137, 35, 7, 1]

73. The distribution of monthly income of 4000 employees follows normal distribution with mean Rs. 6000 and standard deviation Rs. 1000 find :
(*i*) Number of employees having income more than Rs. 7,000;
(*ii*) The number of employees having income less than Rs. 5500;
(*iii*) The least monthly income among the highest paid 100 employees.

74. (*a*) Proof reading of 200 pages of a book containing 500 pages gave the following results :

No. of mistakes per page :	0	1	2	3	4	5
Frequency :	113	62	20	3	1	1
Cost per page for checking :	1.0	1.5	2.5	3.0	3.5	4.0

(*a*) Fit a Poisson distribution.
(*b*) Estimate the total cost of correcting the whole book.
[(*a*) 109.76, 65.85, 19.15, 3.95, 0.592, 0.0711 (*b*) 272.139]

75. Which probability distribution is most likely the appropriate one to use for the following data : Binomial, Poisson or Normal ?
(*i*) The life span of a female born in 1957.
(*ii*) The number of autos passing through a toll booth.
(*iii*) The number of defective radios in a lot of 100.
(*iv*) The water level in a reservoir.

76. A book has 700 pages. The number of pages with various number of misprints is recorded below. Fit a Poisson distribution to the given data :

Number of Misprints X :	0	1	2	3	4	5
Number of Pages with X misprints :	616	70	10	2	1	1

(M.Com., DU, 1999)

77. An insurance salesman sells policies to 5 men, all of an identical age and in good health. According to actuarial tables, the probability that a man of this particular age will be alive 30 years hence is 2/3. Find the probability that 30 years hence :

(*i*) at least 1 man will be alive.

(*ii*) at least 3 men will be alive. *(MBA, IGNOU, 2002)*

78. (*a*) The local authorities in a certain city install 10,000 electric lamps in the streets of the city. If these lamps have an average life of 1000 burning hours with a standard deviation of 200 hours, assuming normality, what number of lamps might be expected to fail in the first 800 burning hours ? *(MBA, IGNOU, 2004)*

(*b*) In certain organisation out of 400 employees 150 are married. Find the probability that exactly 2 of the 3 randomly chosen employees are unmarried. The purchase department has 10 employees. Find the probability that exactly 4 employees of the department are married. *(MBA, Bharathidasan Univ., April 2003)*

79. In an examination, it is laid down that a student passes if he secures 30% 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 a distinction in case he secures 80% or more marks. It is noticed from the results that 10% of the students failed in the examination, whereas 5% of them obtained distinction. Calculate the percentage of the students placed in second division (Assume normal distribution). *(MBA, IGNOU, 2003)*

80. A T.V. manufacturer is facing the problem of selecting a supplier of cathode ray tube, which is the most vital component of a T.V. set. Three foreign suppliers, all equally dependable have agreed to supply the tubes. The prices per tube and the expected life of a tube for the three suppliers are as follows :

	Price/Tube	*Expected life of tube*
Supplier 1	Rs. 800	1500 hrs.
Supplier 2	Rs. 1000	2000 hrs.
Supplier 3	Rs. 1500	4000 hrs.

The manufacturer guarantees its customers that it will replace the T.V. set if the tube fails earlier than 1000 hours. Such a replacement would cost him Rs. 1000/tube, over and above the price of the tube. Can you help the manufacturer to select a supplier ? *(MBA, IGNOU, 2006)*

81. (*a*) A duplicating machine maintained for office use is operated by an office assistant who earns Rs. 50 per hour. The time to complete each job varies according to an exponential distribution with mean 6 minutes. Assume a Poisson input with an average arrival rate of 5 jobs per hour. If an 8 hours day is used as a base, determine the percentage idle time of the machine and the average time a job is in the system. *(B.E./B.Tech., Madras Univ., 2003)*

(*b*) In a survey with a sample of 300 respondents, the monthly income of the respondents follows normal distribution with its mean and standard deviation as Rs. 15,000 and Rs. 3,000 respectively. Answer the following :

(*i*) What is the probability that the monthly income is less than Rs. 12,000 ? Also, find the number of respondent having income less than Rs. 12,000.

(*ii*) What is the probability that the monthly income is more than Rs. 16,000 ? Also, find the number of respondents having income more than Rs. 16,000.

(*iii*) What is the probability that the monthly income, is in between Rs. 10,000 and Rs. 17,000 ? Also, find the number of respondents having income in between Rs. 10,000 and Rs. 17,000. *(MBA, Bharathidasan Univ., 2005)*

(*c*) The mean weight of a lunch rice pack is 0.255 kg with a standard deviation of 0.025. The random variable weights of the pack follows a normal distribution.

(*i*) What is the probability that pack weighs less than 0.280 Kg. ?

(*ii*) What is the probability that pack weighs more than 0.350 Kg. ?

(*iii*) What is the probability that the pack weighs between 0.260 Kg. and 0.340 Kg. ? *(M.A. Eco., M.K. Univ., 2007)*

*****

13

Sampling and Sampling Distributions

INTRODUCTION

The need for adequate and reliable data is ever increasing for taking wise decisions in different fields of human activity and business is no exception to it. There are two ways in which the required information may be obtained :

1. Complete enumeration survey or census method, and
2. Sampling method.

Under complete enumeration survey method, data are collected for each and every unit (person, household, field, shop, factory, etc., as the case may be) belonging to the population or universe which is the complete set of items which are of interest in any particular situation. For example, we are interested in knowing consumers' reactions to a particular product. We may contact each and every person who uses that product or just take a sample of person. In the former case, we are using *census method* while in the latter *sample method*.

The main advantages of census method are :

1. *Information can be obtained for each and every unit.* If the information is required for each and every unit in the domain of study, a complete enumeration survey is necessary. Data for every individual or unit are necessary in cases where the action is taken separately for each one of them. Examples of such situations are recruitment of personnel in an establishment, preparation of voters' list for election purposes, income tax assessment, where the income of each individual is assessed and taxed.

2. *Greater Accuracy.* The results of a complete enumeration survey are expected to be more accurate than the sample method because information is obtained for each and every unit. However, it may be noted that a complete enumeration survey need not necessarily provide us with accurate information as evidenced by the census experience of a number of countries. The errors in a complete enumeration survey arise mainly from incomplete coverage, observational and tabulation errors due to the difficulties encountered in organising a survey on such a large scale and in getting adequate trained personnel to carry out the survey.

The effort, money and time required for carrying out complete enumeration will generally be extremely large, and, in many cases, cost may be so prohibitive that the very idea of collecting information by this method may have to be dropped. The choice, then, may be either no data or data through the sampling method. Unless the information is required for each and every unit in the domain of study, the sampling technique is generally used to obtain the information.

In the sampling method, instead of every unit of the population only a part of the population is studied and the conclusions are drawn on that basis for the entire population.

Although much of the development in the theory of sampling has taken place only in recent years, the idea of sampling is pretty old. A housewife examines only two or three grains of boiling rice to know, whether the pot of rice is ready or not. A doctor examines a few drops of blood and draws

conclusion about the blood constitution of the whole body. A businessman places orders for materials by examining only a small smaple of the same. A teacher may put questions to one or two students and find out whether the class as a whole is following the lesson. In fact, there is hardly any field where the technique of sampling is not used either consciously or unconsciously.

Purpose of Sampling. A sample is not studied for its own sake. The basic objective of its study is to draw inference about the population. In other words, sampling is only a tool which helps to know the characteristics of the universe or population by examining only a small part of it. The values obtained from the study of sample, such as the average and variance, are known as *'statistic'*. On the other hand, such values for the population are called *'parameters'*.

Principles of Sampling

There are two important principles, on which the theory of sampling is based:

1. Principle of 'Statistical Regularity', and
2. Principle of 'Inertia of Large Numbers'.

Principle of Statistical Regularity

This principle is derived from the mathematical theory of probability. In the words of King, *"The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group"*. In other words, this principle points out that if a sample is taken at random from a population, it is likely to possess almost the same characteristics as that of the population. This principle directs our attention to one very important point, that is, the desirability of choosing the sample at random.

By random selection, we mean a selection where each and every item of the population has an **equal** chance of being selected in the sample. In other words, the selection must not be made by deliberate exercise of one's discretion. A sample selected in this manner would be representative of the population. If this condition is satisfied, it is possible for one to depict fairly, accurately the characteristics of the population by studying only a part of it. Hence, this principle is of great practical significance because it makes possible a considerable reduction of the work necessary before any conclusion is drawn regarding a large population. For example, if one intends to make a study of the cigarette buying habits of the students of Delhi University, it is not necessary to study each and every student, a few students may be selected at random from every college, and on that basis inference may be drawn for all the students of the University.

It should be noted that the results derived from sample data may be different from that of the population. This is for simple reason that the sample is only a part of the whole population.

Principle of Inertia of Large Numbers

This principle is a corollary of the principle of statistical regularity. It is of great significance in the theory of sampling. It states that, other things being equal, larger the size of the sample, more accurate the results are likely to be. This is because large numbers are more stable as compared to small ones. The difference in the aggregate result is likely to be significant when the number in the sample is large, because when large numbers are considered, the variations in the component parts tend to balance each other and, therefore, the variation in the aggregate is insignificant. For example, if a coin is tossed 10 times, we should expect an equal number of heads and tails, *i.e.*, 5 each. But since the experiment is tried a small number of times, it is likely that we may not get exactly 5 heads and 5 tails. The result may be combination of 9 heads and 1 tail, or 8 heads and 2 tails, or 7 heads and 3 tails etc. If the same experiment is carried out 1,000 times, the chance of getting 500 heads and 500 tails would be very high, *i.e.*, the results would be very near to 50% heads and 50% tails. The basic reason for such likelihood is that the

experiment has been carried out a sufficiently large number of times and possibility of variations in one direction compensating for others in a different direction is greater. If at one time we get continuously 5 heads, it is likely that at other times we may get continuously 5 tails, and so on, and for the experiment as a whole, the number of heads and tails may be more or less equal. Similarly, if it is intended to study the variation in the production of rice over a number of years and data are collected from one or two States only, the result would reflect large variation in production due to the favourable or unfavourable factors in operation. If,on the other hand, figures of production are collected for all the States in India, it is quite likely that we find little variation in the aggregate. This does not mean that the production would remain constant for all years. It only implies that the changes in the production of the individual States will be counterbalanced and reflect small variation in production for the country as a whole.

Methods of Sampling

When a sample is required to be reflected from a population, it is necessary to decide which method should be applied. The various methods of sampling or sampling designs can be grouped under the heads as random sampling and non-random sampling. Random sampling is also referred to as probability sampling, since, if the sampling process is random, the laws of probability can be applied; thus the pattern of sampling distribution needed to interpret and evaluate a sample is provided. A non-random sample is selected on a basis other than probability considerations such as expert judgment, convenience or some other criteria. The most important aspect of non-random sampling worth noting is that it is subjected to sampling variability but there is no way of knowing the pattern of variability in the process.

We shall now discuss some of the various sampling methods under two separate headings as follows:

A. Random sampling methods :
- (*i*) Simple Random Sampling
- (*ii*) Stratified Sampling
- (*iii*) Systematic Sampling
- (*iv*) Multi-stage Sampling

B. Non-random sampling methods :
- (*i*) Judgment Sampling
- (*ii*) Quota Sampling
- (*iii*) Convenience Sampling.

A. RANDOM SAMPLING METHODS

I. Simple Random Sampling*

Simple random sampling refers to the sampling technique in which each and every item of the population is given an equal chance of being included in the sample. The selection is thus free from personal bias because the investigator does not exercise his discretion of preference in the choice of items. Since selection of items in the sample depends entirely on chance, this method is also known as the method of chance selection. Some people believe that randomness of selection can be achieved by unsystematic and haphazard procedures. But this is quite wrong. However, the point to be emphasized is that unless precaution is taken to avoid bias and a conscious effort is made to ensure the operation of chance factors, the resulting sample shall not to be a random sample.

Random sampling is sometimes referred to as 'representative sampling'. If the sample is chosen at random and if the size of the sample is sufficiently large, it will represent all groups in the population. A

*Simple random samples are characterised by the way in which they are selected. Since, a simple random sample is drawn by chance selection, it must be differentiated from selection in a haphazard or hit-and-miss manner.

random sample is also known as a 'probability sample' because every item of the population has an equal opportunity of being selected in the sample.

Random sampling is not always used as primary sampling procedure. However, it is necessary to introduce an element of randomness in the final selection of items. For example, with each group the choice of cases to constitute the sample should be based on chance selection. If the element of randomness is not introduced, bias is likely to enter and make the sample unrepresentative.

Methods of Obtaining a Simple Random Sample

To ensure randomness of selection, one may adopt any of the following methods:

1. Lottery Method. This is a very popular method of taking a random sample. Under this method, all items of the population are numbered or named on separate slips of paper of identical size,colour and shape. These slips are then folded and mixed up in a container or drum. A blindfold selection is then made of the number of slips required to constitute the desired size of sample. The selection of items thus depends entirely on chance. The method would be quite clear with the help of an example. If we want to take a sample of 10 persons out of a population of 100, the producer is to write the names of all the 100 persons on separate slips of paper, fold these slips, mix them thoroughly and then make a blindfold selection of 10 slips.

The above method is very popular in lottery drawn, where a decision about prizes is to be made. However, while adopting lottery method, it is absolutely essential to see that the slips are of identical size, shape and colour, otherwise there is a lot of possibility of personal prejudice and bias affecting the results.

2. Table of Random Numbers*. The lottery method discussed above becomes quite cumbersome to use as the size of population increases. An alternative method of random selection is that of using the table of random numbers. A number of random number tables are available such as: (*i*) Tippett's table of random numbers, (*ii*) Fisher and Yates numbers, and (*iii*) Kendall and Babington Smith numbers. Tippett's numbers are most popular. They consist of 41,600 digits taken from census reports and combined by fours to give 10,400 four-figure numbers. We give here the first forty sets as an illustration of the general appearance of random numbers :

2952	6641	3992	9792	7969	5911	3170	5624	4167	9524
1542	1396	7203	5356	1300	2693	2370	7483	3408	2792
3563	1089	6913	7691	0560	5246	1112	6107	6008	8126
4233	8776	2754	9143	1405	9025	7002	6111	8816	6446

It is important that the starting point in the table of random numbers be selected in some random fashion so that every unit has an equal chance of being selected.

One may question, and quite rightly, as to how it is ensured that these digits are random. It may be pointed out that the digits in the table were chosen haphazardly, but the real guarantee of their randomness lies in practical complication. Tippett's table of random numbers may be used which is given below. Suppose we have to select 20 items out of 6,000. The procedure is to number all the 6,000 items from 1 to 6,000. A page from Tippett's table may then be consulted and the first twenty numbers up to 6,000 noted down. Items bearing these numbers will be included in the sample. Making use of the portion of the table, given above, the required numbers are :

2952	3992	5911	3170	5624	4167	1542	1396	5356	1300
2693	2370	3408	2762	3563	1089	0560	5246	1112	4233

The items which bear the above numbers constitute the sample.

*These days, Computerised Random Numbers are more popular in use.

Population size less than 1,000. If the size of population is less than 1,000, the procedure will be different, as Tippett's numbers are available only in four figures. Thus, for example, if it is desired to take a sample of 10 items out of 400, all items from 1 to 400 should be numbered as 0001 to 0400. We may now select 10 numbers from the table which are up to 0400.

Population size less than 100. If the size of population is less than 100, the table is used as follows: Suppose ten numbers from out of 0 to 80 are required. We start anywhere in the table and write down the numbers in pairs. The table can be read horizontally, vertically, diagonally, or in any other methodical way. Starting with the first and reading horizontally (please see the part of table given on page 461), we obtain 29, 52, 66, 41, 39, 92, 97, 92, 79, 69, 59, 11, 31, 70, 56, 24, 41, 67, and so on. Ignoring the numbers greater than 80, we obtain for our purpose ten random numbers, namely, 29, 52, 66, 41, 39, 79, 69, 59, 11 and 31.

Fisher and Yates tables consist of 15,000 numbers. These have been arranged in two digits in 300 blocks, each block consisting of 5 rows and 5 columns. Kendall and Smith also constructed random numbers (10,000 in all) by using a randomising machine. However, this method of random selection cannot be followed in case of articles like ghee, oil, petrol, wheat, etc.

Merits. 1. Since the selection of items in the sample depends entirely on chance, there is no possibility of personal bias affecting the results.

2. As compared to judgment sampling, random sample represents the population in a better way. As the size of the sample increases, it becomes increasingly representative of the population.

3. The analyst can easily assess the accuracy of his estimate because sampling errors follow the principle of chance. The theory of random sampling is further developed than that of any other type of sampling which enables the surveyor to provide the most reliable information at the least cost.

Limitations. 1. The use of random sampling necessitates a completely catalogued population from which to draw the sample. But, it is often difficult for the investigator to have up-to-date lists of all the items of the population to be sampled. This restricts the use of any sampling method.

2. The task of preparing slips is time-consuming and expensive. However, this difficulty can at times be overcome by following regular interval sampling method which enable a random sample to be drawn without preparing slips.

3. The size of the sample required to ensure statistical reliability is usually large under random sampling than in stratified sampling.

4. From the point of view of field survey, it has been claimed that cases selected by random sampling tend to be too widely dispersed geographically and that the time and cost of collecting data become too large.

5. Random sampling may produce the most non-random looking results. For example, thirteen cards from a well-shuffled pack of playing cards may consist of one suit. But the probability of this type of incidence is very-very small.

II. Stratified Sampling

Stratified random sampling is one of the restricted random methods which, by using available information concerning the data attempts to design a more efficient sample than that obtained by the simple random procedure. The process of stratification requires that the population may be divided into homogeneous groups or classes called *strata.* Then a sample may be taken from each group by simple random method, and the resulting sample is called a stratified sample.

A stratified sample may be either proportional or disproportionate. In a proportional stratified sampling plan, the number of items drawn from each stratum is proportional to the size of the strata. For instance, if the population is divided into four strata, their respective sizes being 15, 10, 20, 55 per cent of the population and a sample of 1,000 is to be drawn, the desired proportional sample may be obtained in the following manner:

From stratum	one	1,000 (0.15)	= 150 items
"	" two	1,000 (0.10)	= 100 "
"	" three	1,000 (0.20)	= 200 "
"	" four	1,000 (0.55)	= 550 "
	Sample size		=1,000

Proportional stratification yields a sample that represents the population with respect to the proportion in each stratum in the population. This procedure is satisfactory if there is no great difference in variation from stratum to stratum. But, it is certainly not the most efficient procedure, especially when there is considerable variation in different strata. This indicates that in order to obtain maximum efficiency in stratification, we should assign greater representation to a stratum with a large variation and smaller representation to one with small variation. For instance, in conducting an all-India survey of the market of a new product, all the States of India may be taken as strata. If the potential consumers of a given State are 10 per cent of all the consumers, but according to our information the product in that State is bound to have a market, we may take, say, 1 per cent or 2 per cent of our sample from that State. If, however, the outcome of another State is highly doubtful we may decide to give it a much greater representation in the sample from its relative size. A sample, thus obtained is a disproportionate stratified sample. Disproportionate stratified sampling also includes procedures of taking an equal number of items from each stratum irrespective of its size.

Merits. 1. Since the population is first divided into various strata and then a sample is drawn from each stratum there is little possibility of any essential group of the population being completely excluded. A more representative sample is thus secured. Stratified sampling is frequently regarded as the most efficient system of sampling.

2. Stratified sampling ensures greater accuracy. The accuracy is maximum if each stratum is so formed that it consists of uniform or homogeneous items.

3. As compared to random sample, stratified samples can be more concentrated geographically. Thus, the time and expense of interviewing may be considerably reduced.

Limitations. 1. Utmost care must be exercised in dividing the population into various strata. Each stratum must contain, as far as possible, homogeneous items as otherwise the results may not be reliable. However, this is a very difficult task and may involve considerable time and expense.

2. The items from each stratum should be selected at random. But, this may be difficult to achieve in the absence of skilled sampling supervisors and a random selection within each stratum may not be ensured.

III. Systematic Sampling

This method is popularly used in those cases where a complete list of the population from which sampling is to be drawn is available. The method is to select every *k*th* item from the list where '*k*' refers to the sampling interval. The starting point between the first and the *k*th item is selected at random.

$$*k = \frac{\text{Size of Population}}{\text{Sample size}} = \frac{N}{n}$$

For example, if a complete list of 1,000 students of a college is available and if we want to draw a sample of 200 students, this means we must take every fifth item (*i.e.*, $k = 5$). The first item between one and five shall be selected at random. Suppose it comes out to be three. Now, we shall go on adding five and obtain numbers of the desired sample. Thus, the second item would be the 8th student; the third, 13th student; the fourth, 18th student; and so on.

Systematic sampling is relatively a simple technique and may be more efficient than simple random sampling, provided the lists are arranged wholly as random. However, it is rare that this requirement is fulfilled. The nearest approach to randomness is provided by alphabetical lists such as are found in telephone directory, although even these may have certain non-random characteristics.

Merits. The systematic sampling is more convenient to adopt than the random sampling or the stratified sampling method. The time and work involved in sampling by this method are relatively smaller. The results obtained are also found to be generaly satisfactory provided care is taken to see that there are no periodic features associated with the sampling interval. If populations are sufficiently large, systematic sampling can often be expected to yield results that are similar to those obtained by proportional stratified sampling.

Limitations. Systematic sampling becomes a less representative design than simple random sampling if we are dealing with populations having hidden periodicities. For example, if the sales of every seventh day of the calendar year are included, the sample will contain, say, all Mondays or all Fridays. If there is a definite repetitive weekly pattern in sales (which is usually the case), our sample is not representative at all of sales for the whole year and consequently, the sample results may be seriously biased.

IV. Multi-stage Sampling

As the name implies, this method refers to a sampling procedure which is carried out in several stages. The material is regarded as made up of a number of first stage sampling units, each of which is made of a number of second stage units, etc. At first, the first stage units are sampled by some suitable method, such as random sampling. Then, a sample of second stage units is selected from each of the selected first stage units again by some suitable method which may be the same or different from the method employed for the first stage units. Further stages may be added as required. The procedure may be illustrated as follows:

Suppose, we want to take a sample of 5,000 households from the State of U.P. At the first stage, the State may be divided into a number of districts and a few districts selected at random. At the second stage, each district may be sub-divided into a number of villages and a sample of villages may be taken at random. At the third stage, a number of households may be selected from each of the villages selected at the second stage. In this way, at each stage, the sample size becomes smaller and smaller.

Merits. Multi-stage sampling introduces flexibility in the sampling method which is lacking in other methods. It enables existing divisions and sub-divisions of the population to be used as units at various stages, and permits the field work to be concentrated and yet large area to be covered. Another advantage of the method is that sub-division into second stage unit (*i.e.*, the construction of the second stage frame) need be carried out for only those first stage units which are included in the sample. It is, therefore, particularly valuable in surveys of underdeveloped areas where no frame is generally sufficiently detailed and accurate for sub-division of the material into reasonably small sampling units.

Limitations. However, a multi-stage sample is in general less accurate than a sample containing the same number of final stage units which have been selected by some suitable stage process.

B. NON-RANDOM SAMPLING METHODS

Random selection is generally recommended for large surveys, but certain types of non-random selection are sometimes justified. A few of the most important of these types are:

I. Judgment Sampling

In this method of sampling, the choice of sample items depends exclusively on the judgment of the investigator. In other words, the investigator exercises his judgment in the choice of sample items and includes those items in the sample which he thinks are most typical of the population with regard to the characteristics under investigation. For example, if a sample of ten students is to be selected from a class of sixty for analysing the spending habits of students, the investigator would select 10 students who, in his opinion, represent the class.

This method, though simple, is not scientific because the results may be considerably affected by the personal prejudice or bias of the investigator. Thus, judgment sampling involves the risk that the investigator may establish foregone conclusions by including those items in the sample which conform to his preconceived notions. For example, if an investigator holds the view that the wages of workers in a certain establishment are very low, and if he adopts the judgment sampling method, he may include only those workers in the sample whose wages are low and thereby establish his point of view which may be far from the truth.

Even though the principles of sampling theory are not applicable to judgment sampling, this method is often used in solving many types of economic and business problems such as:

(*i*) Judgment sampling is used when size of sample is small. In such a case, simple random sample may miss the more important elements, whereas judgment selection would certainly include them in the sample.

(*ii*) In solving everyday business problems and making public policy decisions, executives and public officials are often pressed for time and cannot wait for probability sample designs. Judgment sampling is then the only practical method, since estimates can be made available quickly that will enable businessmen and governmental officials to arrive at solutions to their urgent problems that are better than decisions made without any statistical data.

(*iii*) Judgment sampling may be used to conduct pilot studies. In any case, the reliability of sample results in judgment sampling depends on the quality of the sampler's expert knowledge or judgment. If it is good and is carefully and skilfully applied, judgment samples may be expected to be representative and to yield valuable results. On the other hand, when a sample is obtained with poor judgment, serious bias will be present.

The success of this method depends upon the excellence in judgment. If the individual making decisions is knowledgeable about the population and has good judgment, then the resulting sample may be representative, otherwise the inferences based on the sample may be erroneous. It may be noted that even if a judgment is reasonably representative, there is no objective method for determining the size or likelihood of sampling error. This is a big defect of the method.

II. Quota Sampling

Quota sampling is a type of judgment sampling. In a quota sample, quotas are set up according to given criteria, but, within the quotas the selection of sample items depends on personal judgment. For example, in a radio listening survey, the interviewers may be told to interview 500 people living in a certain area and that out of every 100 persons interviewed 60 are to be housewives, 25 farmers and 15 children under the age of 15. Within these quotas, the interviewer is free to select the people interviewed.

The cost per person interviewed may be relatively small for a quota sample but there are numerous opportunities for biases which may invalidate the results. For example, interviewers may miss farmers working in the fields or talk with those housewives who are at home. If a person refuses to respond, the interviewer simply selects someone else. Because of the risk of personal prejudice and bias entering the process of selection, the quota sampling is rarely used in practical work.

III. Convenience Sampling

The method of convenience sampling is also called the chunk. A chunk is a fraction of one population taken for investigation because of its convenient availability. Thus, a chunk is selected neither by probability nor by judgment but by convenience. A sample obtained from readily available lists, such as telephone directories or automobile registrations, is a convenience sample and not a random sample, even if the sample is drawn at random from the lists.

Convenience samples are sometimes called accidental samples because those entering into the sample enter by "accident"—they just happen to be at the right place and at the right time, that is, where and when the information for the study is being collected. The problem with convenience samples is that we have no way of knowing if those included in the sample are representatives of the target population.

A chunk—which is merely a convenient slice of the population—can hardly be representative of the population. Its results are generally biased and unsatisfactory. Formerly, the chunk was frequently used in public opinion surveys when interviewers stopped near the railway station or the bus stop or in front of office building to interview people. Today, accountants still use convenience sampling to analyse or audit accounts.

Convenience sampling is also useful in making pilot studies. Questions may be tested and preliminary information may be obtained by the chunk before the final sampling design is decided upon.

Size of Sample

An important decision that has to be taken while adopting a sampling technique is about the size of the sample. Different opinions have been expressed by experts on this point. For example, some have suggested that the sample size should be 5% of the size of population while others are of the opinion that sample size should be at least 10%. However, these views are of little use, as in practice, appropriate sample size depends on various factors relating to the subject under investigation like the time aspect, the cost aspect, the degree of accuracy desired, etc. Sampling theory is of little help in arriving at a good estimate of the sample size in any particular situation. However, the following two considerations may be kept in mind in determining the appropriate size of the sample.

1. The size of the sample should increase as the variation in the individual items increases.
2. The greater the degree of accuracy desired, the larger should be the sample size.

Merits of Sampling Method

The sampling method has the following merits over the complete enumeration survey method :

1. *Less time.* Since the sample is a study of part of the population, considerable time and labour are saved when a sample survey is carried out. Time is saved not only in collecting data but also in processing it. For these reasons, a sample provides more timely data in practice than a census.

2. *Less cost.* The amount of effort and expenses involved in collecting information is always greater per unit of the sample than a complete census, the total financial burden of a sample survey is generally less than that of a complete census. This is because of the fact that in sampling, we study only a part of the population and the total expense of collecting data is less than that required when the census method is adopted. This is a great advantage particularly in an underdeveloped economy where much of the information would be difficult to collect by the census method for lack of adequate resources.

3. *More reliable results.* Although the sampling technique involves certain inaccuracies owing to sampling errors, the result obtained is generally more reliable than that obtained from a complete count. There are several reasons for it. *First,* it is always possible to determine the extent of sampling errors. *Secondly,* other types of errors to which a survey is subjected, such as inaccuracy of information, incompleteness of returns, etc., are likely to be more serious in a complete census than in a sample survey. This is because more effective precautions can be taken in a sample survey to ensure that the information is accurate and complete. Moreover, it is possible to avail of the services of experts and to impart thorough training to the investigators in sample survey which further reduces the possibility of errors. Follow-up work can also be undertaken much effectively in the sampling method. Indeed, even a complete census can only be tested for accuracy by some type of sampling check.

4. *More detailed information.* Since the sampling technique saves time and money, it is possible to collect more detailed information in a sample survey. For example, if the population consists of 1,000 persons in a survey of the consumption pattern of the people, the two alternative techniques available are as follows :

(*a*) We may collect the necessary data from each one of the 1,000 persons through a questionnaire containing, say, 10 questions (census method), or

(*b*) We may take sample of 100 persons (*i.e.*, 10% of population) and prepare a questionnaire containing as many as 100 questions. The expenses involved in the latter case would almost be the same as in the former, but it will enable many times more information to be obtained.

5. *The destructive nature of certain tests.* Many tests are of destructive nature. Steel plates, wires and other similar products often must have a certain minimum tensile strength. To ensure that the product meets the minimum standard, a relatively small sample is selected. Each piece is stretched until it breaks, and the breaking point (usually measured in pounds) is recorded. Obviously, if all the wires or all the plates were tested for tensile strength, none would be available for sale or use. And for the same reason, only a sample of photographic film is selected to determine the quality of all the film produced, only a few seeds are tested for germination prior to planting season and only a few chalks out of a certain lot are tested for ascertaining the breaking strength.

Limitations of Sampling

Despite the various advantage of sampling, it is not altogether free from limitations. Some of the precautions involved in sampling are given below :

(*i*) A sample survey must be carefully planned and executed, otherwise, the results obtained may be inaccurate and misleading. Of course, even for a complete count care must be taken but serious errors may arise in sampling, if the sampling procedure is not perfect.

(*ii*) Sampling generally requires the services of experts, if only for consultation purposes. In the absence of qualified and experienced persons, the information obtained from sample surveys cannot be relied upon. A shortage of experts in the sampling field is a serious hurdle in the way of reliable statistics.

(*iii*) At times, the sampling plan may be so complicated that it requires more time, labour and money than a complete count. This is so if the size of the sample is a large proportion of the total population and if complicated weighted procedures are used. With each additional complication in the survey, the chances of errors multiply and greater care has to be taken which, in turn, means more time and labour.

(*iv*) If the information is required for each and every unit in the domain of study, a complete enumeration survey is necessary.

Sampling and Non-Sampling Errors

The term 'error' refers to the difference between the value of a 'sample statistic' and that of corresponding 'population parameter'. Various forces combine to produce deviations of sample statistic from population parameters, and errors, in accordance with the different causes, are classified into sampling and non-sampling errors.

The error arising due to drawing inferences about the population on the basis of few observations (sample) is termed as sampling error. Clearly, the sampling error in this sense is non-existent in a complete enumeration survey, since the whole population is surveyed. However, the errors mainly arising at the stages of ascertainment and processing of data which are termed non-sampling errors are common both in complete enumeration and sample surveys.

I. Sampling Errors

Even if utmost care has been taken in selecting a sample, the results derived from the sample may not be representative of the population from which it is drawn, because samples are seldom, if ever, perfect miniatures of the population. This gives rise to sampling errors. Sampling errors arise due to the fact that samples are used and to the particular method used in selecting the items from the population.

Sampling errors are of two types—biased and unbiased.

(1) *Biased errors.* These errors arise from any bias* in selection, estimation, etc. For example, if in place of simple random sampling, deliberate sampling has been used in a particular case, some bias is introduced in the result and hence such errors are called biased sampling errors.

(2) *Unbiased errors.* These errors arise due to chance differences between the members of population included in the sample and those not included.

Thus, the total sampling errors is made up of errors due to bias, if any, and the random sampling error. The essence of bias is that it forms a constant component of error that does not decrease in a large population as the number in the sample increases. Such error is, therefore, also known as *cumulative* or *non-compensating error*. The random sampling error, on the other hand, decreases on an average as the size of the sample increases. Such error is, therefore, also known as non-cumulative of compensating error.

Causes of Bias

Bias may arise due to:

(1) faulty process of selection;

(2) faulty work during the collection of information; and

(3) faulty methods of analysis.

(1) Faulty Selection. Faulty selection of the sample may give rise to bias in a number of ways, such as :

(*a*) *Deliberate selection* of a 'representative' sample.

(*b*) Conscious or unconscious *bias in the selection of a 'random' sample.* The randomness of selection may not really exist even though the investigator claims that he has a random sample if he allows his desire to obtain a certain result to influence his selection.

(*c*) *Substitution.* Substitution of an item in place of one chosen in a random sample sometimes leads to bias. Thus, if it were decided to interview every 50th householder in the street, it would be inappropriate to interview the 51st or any other number in his place as the characteristics possessed by them may differ from those who were originally to be included in the sample.

*Bias is said to exist when the value of a sample statistic shows a persistent tendency to deviate in one direction from the value of the parameter.

(*d*) *Non-response.* If all the items to be included in the sample are not covered, there will be bias even though no substitution has been attempted. This fault particularly occurs in mailed questionnaires, which are incompletely returned. Moreover, the information supplied by the informants may also be biased.

(*e*) *An appeal to the vanity* of the person questioned may give rise to yet another kind of bias. For example, the question 'Are you a good student ?' is such that most of the students would succumb to vanity and answer 'Yes'.

(2) Bias due to Faulty Collection of Data. Any consistent error in measurement will give rise to bias whether the measurements are carried out on a sample or on all the units of the population. The danger of error is, however, likely to be greater in sampling work, since the units measured are often smaller. Bias may arise due to improper formulation of the decision problems wrongly defining the population, specifying the wrong decision, securing and inadequate frame, and so on. Biased observations may result from a poorly designed questionnaire, an ill-trained interviewer, failure of a respondent's memory, etc. Bias in the flow of data may be due to unorganised collection procedure, faulty editing or coding of responses.

(3) Bias in Analysis. In addition to bias which arises from faulty process of selection and faulty collection of information, faulty method of analysis may also introduce bias. Such bias can be avoided by adopting the proper methods of analysis.

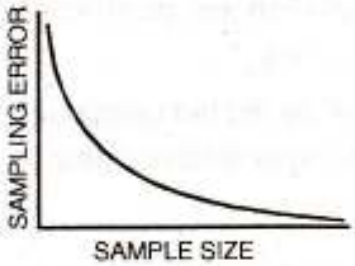
Avoidance of Bias

If possibilities of bias exist, fully objective conclusions cannot be drawn. The first essential of any sampling or census procedure must, therefore, be the elimination of all sources of bias. The simplest and the only certain way of avoiding bias in the selection process is for the sample to be drawn either entirely at random or at random subject to restrictions, which, while improving the accuracy, are of such a nature that they do not introduce bias in the results. In certain cases, systematic selection may also be permissible.

Method of Reducing Sampling Errors

Once the absence of bias has been ensured, attention should be given to the random sampling errors. Such errors must be reduced to the minimum so as to attain the desired accuracy.

Apart from reducing errors of bias, the simplest way of increasing the accuracy of a sample is to increase its size. The sampling error usually decreases with increase in sample size (number of units selected in the samples) and in fact in many situations the decrease is inversely proportional to the square root of the sample size as can be seen from the diagram given below.



From this diagram, it is clear that though the reduction in sampling error is substantial for initial increases in sample size, it becomes marginal after a certain stage. In other words, considerably greater effort is needed after a certain stage to decrease the sampling error than in the initial instances. Hence, after that stage sizable reduction in cost can be achieved by lowering even slightly the precision required. From this point of view, there is a strong case for resorting to a sample survey to provide estimates within permissible margins to error instead of a complete enumeration survey, as in the latter, the effort and the cost needed will be substantially higher due to the attempt to reduce the sampling error to zero.

II. Non-sampling Errors

When a complete enumeration of units in the universe is made, one would expect that it would give rise to data free from errors. However, in practice, it is not so. For example, it is difficult to completely avoid errors of observation or ascertainment. So, also in the processing of data tabulation errors may be committed affecting the final results. Errors arising in this manner are termed *non-sampling errors*, as they are due to factors other than the inductive process of inferring about the population from a sample. Thus, the data obtained in an investigation by a complete enumeration, although free from sampling error, would still be subjected to non-sampling error, whereas the result of a sample survey would be subjected to sampling error as well as non-sampling error.

As regards non-sampling errors, they are likely to be more in case of complete enumeration survey than in case of a sample survey, since it is possible to reduce the non-sampling errors to a greater extent by using better organisation and suitably trained personnel at the field and tabulation stages. The behaviour of the non-sampling error with increase in sample size is likely to be opposite of that of sampling error, that is, the non-sampling error is likely to increase with increase in sample size. In many situations, it is quite possible that the non-sampling error in a complete enumeration survey is greater than both the sampling and non-sampling errors taken together in a sample survey, and naturally in such a situation, the latter is to be preferred to the former.

Non-sampling errors can occur at every stage of planning and execution of the census or survey. Such errors can arise due to a number of causes such as defective methods of data collection and tabulation, faulty definition, incomplete coverage of the population or sample, etc. More specifically, non-sampling errors may arise from one or more of the following factors :

1. Data specification being inadequate and inconsistent with respect to the objectives of the census or survey.

2. Omission or duplication of units due to imprecise definition or boundaries of area units, incomplete or wrong identification of units or faulty methods of enumeration.

3. Inaccurate or inappropriate methods of interview, observation or measurement with inadequate or ambiguous schedules, definitions or instructions.

4. Lack of trained and experienced investigators.

5. Lack of adequate inspection and supervision of primary staff.

6. Errors due to non-response, *i.e.*, incomplete coverage in respect of units.

7. Errors in data processing operations such as coding, punching, certification, tabulation, etc.

8. Errors committed during presentation and printing of tabulated results.

9. Inadequate scrutiny of the basic data.

This should not be taken as a complete list but comprises major sources of error. In a sample survey, non-sampling errors may also arise due to defective frame and faulty selection of sampling units.

Control of Non-sampling Errors

In some situations, the non-sampling errors may be large and deserve greater attention than the sampling errors. While in general, sampling error decreases with increase in sample size, non-sampling error tends to increase with the sample size. In the case of complete enumeration, non-sampling errors and in the case of sample surveys, both sampling and non-sampling errors require to be controlled and reduced to a level at which their presence does not vitiate the use of final results.

In recent years, there has been a growing need for assessing and controlling the non-sampling errors that are likely to arise at the various stages of collection and tabulation of statistical data in

large-scale census and surveys. The increasing awareness of the existence of such errors is due to the fairly widespread use of the sampling method, one of the main advantages of which is that it provides an opportunity for greater control of non-sampling errors as well.

SAMPLING DISTRIBUTIONS

Much of the information used in business and industry is gathered by means of sampling. It has been pointed out earlier that not only it is often impossible either physically or because of limitations imposed by time or pecuniary considerations, to take a census of all the items in the population, but it is also usually unnecessary. The results of a properly taken sample, if subjected to rigorous analysis, will ordinarily enable the investigator to arrive at generalisations that are valid for the entire population.

The process of generalising these sample results of the population is referred to as statistical inference. In this chapter, along with the knowledge of certain probability distributions, we shall use certain sample statistics (such as the sample mean, the sample proportion, etc.) in order to estimate and draw inferences about the true population parameters.

For example, in order to be able to use the sample mean to estimate the population mean, we should examine every possible sample (and its mean) that could have occurred in the process of selecting one sample of a certain size. If this selection of all possible samples actually were to be done, the distribution of the results would be referred to as a sampling distribution. Although, in practice, only one such sample is actually selected, the concept of sampling distributions must be examined so that probability theory and its distribution can be used in making inferences about the population parameter values.

Sampling theory has made it possible to deal effectively with these problems. However, before we discuss in detail about them from the standpoint of sampling theory, it is necessary to understand the Central Limit Theorem and the following three probability distributions, their characteristics and relations :

(1) The population (universe) distribution,

(2) The sample distribution, and

(3) The sampling distribution.

Central Limit Theorem. The Central Limit Theorem, first introduced by De Moivre during the early eighteenth century, happens to be the most important theorem in statistics. According to this theorem, if we select a large number of simple random samples, say, from any population distribution and determine the mean of each sample, the distribution of these sample means will tend to be described by the normal probability distribution with a mean μ and variance σ^2/n. This is true even if the population distribution itself is not normal. Or, in other words, we can say that the sampling distribution of sample means approaches to a normal distribution, irrespective of the distribution of population from where sample is taken and approximation to the normal distribution becomes increasingly close with increase in sample size. Symbolically, the theorem can be explained as follows :

When given n independent random variables $X_1, X_2, X_3, \ldots\ldots X_n$, which have the same distribution (no matter what the distribution), then :

$$X = X_1 + X_2 + X_3 + \ldots + X_n$$

is a normal variate. The mean μ and variance σ^2 of X are

$$\mu = \mu_1 + \mu_2 + \mu_3 + \ldots\ldots + \mu_n = n\mu_i$$

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots\ldots + \sigma_n^2 = n\sigma_i^2$$

where μ_1 and σ_1^2 are the mean and variance of X_i.

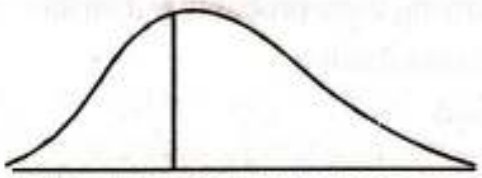
The utility of this theorem is that it requires virtually no conditions on distribution patterns of the individual random variable being summed. As a result, it furnishes a practical method of computing approximate probability values associated with sums of arbitrarily distributed independent random

variables. This theorem helps to explain why a vast number of phenomena show approximately a normal distribution. Consider a case when the population is skewed, skewness of the sampling distribution of means is inversely proportional to the square root of the sample size. Consider the case when $n = 16$ that means the sampling distribution of means will exhibit only one-fourth as much skewness as the population has. Consider the case when $n = 100$, skewness becomes one-tenth as much, *i.e.*, as the sample size increases, the skewness will decrease. As a practical consequence, the normal curve will serve as a satisfactory model when samples are small and population is close to a normal distribution, or when samples are large and population is markedly skewed. Because of its theoretical and practical significance, this theorem is considered as most remarkable theoretical formulation of all probability laws.

The Population (Universe) Distribution

When we talk of population distribution, we assume that we have investigated the population and have full knowledge of its mean and standard deviation. For example, a company might have manufactured 1,00,000 tyres of cars in the year 2004. Suppose, it contacts all those who had bought these tyres and gathers information about the life of these tyres. On the basis of the information obtained, the mean of the population which is also called true mean symbolised by μ and its standard deviation symbolised by σ can be worked out. These Greek letters μ and σ are used for these measures to emphasise their difference from corresponding measures taken from a sample. It may be noted such measures characterising a population are called population *parameters*.

The shape of the distribution of the life of tyres may be as follows :



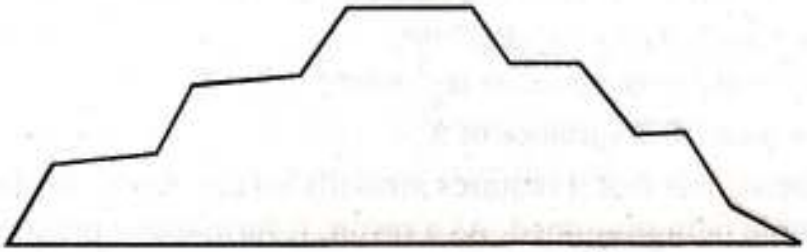
DISTRIBUTION OF THE LIFE OF TYRES

It is clear from above that, though, the distribution shows slight skewness, it does not depart radically from a normal distribution. However, this should not lead one to the conclusion that for sampling theory to apply, it is necessary that the distribution must be normally distributed.

The Sample Distribution

When we talk of a sample distribution, we take a sample from the population. A sample distribution may take any shape. The mean and standard deviation of the sample distribution are symbolised by $\bar{x}$ and s respectively. A measure characterising a sample such as $\bar{x}$ or s is called a sample *statistic*. It may be noted that several sample distributions are possible from a given population.

Suppose, in the above illustration, the manufacturer takes a sample of 500 tyres. He contacts th buyers and enquires about the life of tyres. The shape of the distribution of these tyres may be as follows



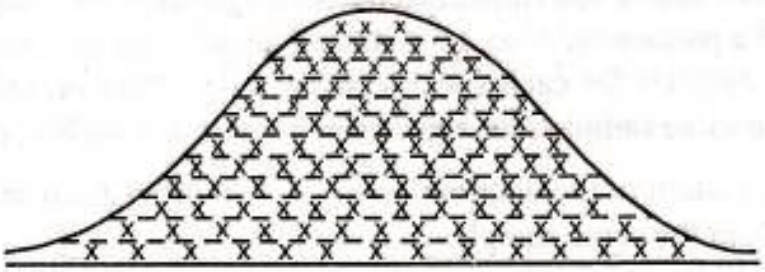
SAMPLE DISTRIBUTION OF 500 TYRES

The mean values of these tyres can be expected to differ somewhat from one sample to another. These sample means constitute the raw material out of which a sampling distribution is constructed.

The Sampling Distribution

Sampling distributions constitute the theoretical basis of statistical inference and are of considerable importance in business decision-making. If we take numerous different samples of equal size from the same population, the probability distribution of all the possible values of a given statistic from all the distinct possible samples of equal size is called a sampling distribution.

It is interesting to note that sampling distributions closely approximate a normal distribution. It can be seen that the mean of a sampling distribution of sample means is the same as the mean of the population distribution from which the samples were taken.* The following diagram would make it clear :



SAMPLING DISTRIBUTION OF SAMPLE MEANS

The mean of the sampling distribution is designated by the same symbol as the mean of the population, namely μ. However, the standard deviation of the sampling distribution of means given a special name, *standard error of mean*, and is symbolised by $\sigma_{\bar{x}}$. The subscript indicates that in this case, we are dealing with a sampling distribution of means.

The greatest importance of sampling distributions is the assistance that they give us in revealing the patterns of sampling errors and their magnitude in terms of standard error. In sampling with replacement, we can observe a good deal of fluctuations in the sample mean as compared to fluctuations in the actual population. The fact that the sample means are less variable than the population data follows logically from an understanding of the averaging process. A particular sample mean averages together all the values in the sample. A population (universe) may consist of individual outcomes that can take on a wide range of values from extremely small to extremely large. However, if an extreme value falls into the sample, although it will have an effect on the mean, the effect will be reduced since it is being averaged in with all the other values in the sample. Moreover, as the sample size increases, the effect of a single extreme value gets even smaller, since it is being averaged with more observations. This phenomenon is expressed statistically in the value of the standard deviation of the sample mean. This is

*Let $x_1, x_2, \ldots\ldots x_n$ represent independent random variables corresponding to the n observations in a sample from a population having the same mean μ.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\bar{x} = E(\bar{x}) = E\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right]$$

$$= \frac{1}{n}\{E[x_1] + E[x_2] + \ldots + E[x_n]\}$$

$$= \frac{1}{n}\{\mu + \mu + \ldots + \mu\} = \frac{1}{n} \cdot n\mu = \mu$$

the measure of variability of the mean from sample to sample and is referrd to as the standard deviation of the sampling distribution of sample mean or *the standard error of the mean* denoted by $\sigma_{\bar{x}}$ and is calculated by*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This formula holds only when population is infinite or samples are from finite population with replacement.

It may be noted that in deducing a sampling distribution, we must first make an assumption about the appropriate population parameter. In as much as any value can be assumed for a parameter, depending upon our knowledge or guess of the population, there is no theoretical limit to the number of sampling distribution for the same sample size that can be taken from the population. There is a sampling distribution for each assumed value of a parameter. Also, given the assumed value of a parameter, there is a different sampling distribution of statistics for each specific sample size. Further, under the same assumptions about a population and the same sample size, the distribution of one statistic differs from that of another statistic. For example, the pattern of the distribution of $\bar{X}$ will differ from that of s^2, even though both measures are computed from the same sample.

Relationship between Population, Sample and Sampling Distributions

It will be interesting to note that the mean of the sampling distribution is the same as the mean of the population. It is possible that many sample means may differ from the population mean. However, the sample information can be used as an estimate of population values.

It has also been established that the observed standard deviation of a sample is close to the standard deviation of the population values.

In fact, the standard deviation of the samples is usually so good an approximation that it can safely be used as an estimate of the corresponding population measure. In order to use s of the sample to estimate σ of the population, we make a slight adjustment which has been found to contribute to greater accuracy of the estimate. The adjustment consists of using $(n-1)$ instead of n in the formula for the standard deviation of a sample, *i.e.*, we use

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \text{ instead of } \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

The adjustment decreases the denominator and, therefore, gives a larger result. Thus, the estimated standard deviation of the population is slightly larger than the observed standard deviation of the sample.

Sampling Distribution of the Mean

If a population distribution is normal, the sampling distribution of the mean $(\bar{x})$ is also normal for samples of all sizes as can be seen from the following diagram :

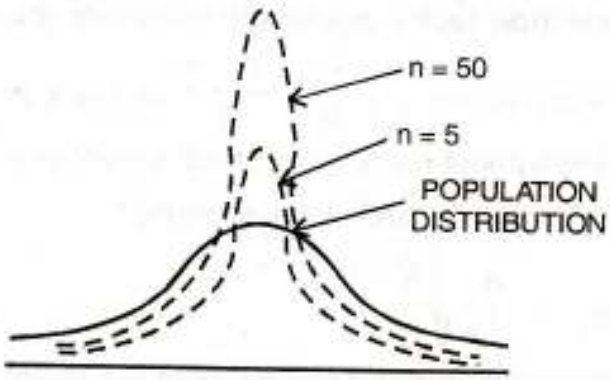
---

*Let $x_1, x_2, \ldots, x_n$ be independent random variables, each having the same variance σ^2.

$$\sigma_{\bar{x}}^2 = \text{Var}(\bar{x}) = \text{Var}\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right]$$

$$= \frac{1}{n^2}[\text{Var}(x_1) + \text{Var}(x_2) + \ldots + \text{Var}(x_n)]$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \text{ or } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



Relationship between a normal population distribution and sampling distribution of the mean for $n = 5$ and $n = 50$.

The following are the important properties of the sampling distribution of mean :

(1) It has a mean equal to the population mean, *i.e.*, $\mu_{\bar{x}} = \mu$.

(2) It has a standard deviation equal to the population standard deviation divided by the square root of the sample size. That is :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{x}}$ is a measure of the spread of $\bar{x}$ values around μ or a measure of average sampling error or simply stated *standard error of the mean.*

σ = standard deviation of the population

n = size of the sample.

(3) It is normally distributed. The distribution of sample means for large samples is distributed normally whatever the shape of the population distribution, provided σ is finite. Samples of 30 or more items are frequently considered large for statistical purposes. It may be pointed out that, if a population is normal, the distribution of sample means is normal, even if the sample size is small.

It should be noted that $\sigma_{\bar{x}}$ is a measure of the precision with which the sample mean can be used to estimate the true population mean, μ, the standard error, $\sigma_{\bar{x}}$ varies directly with the variation in the original population, σ, and inversely with the square root of the sample size n. Thus, as might be expected, the greater the variation among the items in the original population, the greater is the expected sampling error in using $\bar{x}$ as an estimate of μ. Also the larger the sample size, the smaller the standard error and the smaller the sample size, the larger the standard error.

In practice, the standard deviation of the population is rarely known, and therefore, the standard deviation of the samples which closely approximates the standard deviation of the population is used in place of σ. Hence, the formula for standard error takes the following form :

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where s refers to the standard deviation of the sample.

The central limit theorem and the standard error of a sample statistic were based upon the premise that the samples selected were chosen with replacement. However, in survey research and in business, sampling is conducted without replacement from populations that are of a finite size N. In these cases, particularly, when the sample size n is not small as compared to the population size N, a *finite population correction factor* should be used in developing the particular sampling distribution.

The finite population correction factor essentially expresses the proportion of observations th have not been included in the sample, *viz.*, $1 - \frac{n}{N} = \frac{N-n}{N}$ and is approximately equal to $\frac{N-n}{N-1}$ whe N is large. Therefore, in sampling without replacement from a finite population, the sampling distributic of a sample mean will have mean $\mu_{\bar{x}} = \mu$ and standard error,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Illustration 1. The time between two arrivals in a queuing model is normally distributed with a mean 2 minutes a standard deviation 0.25 minute. If a random sample of size 36 is drawn, what is the probability that the sample mean will greater than 2.1 minutes ?

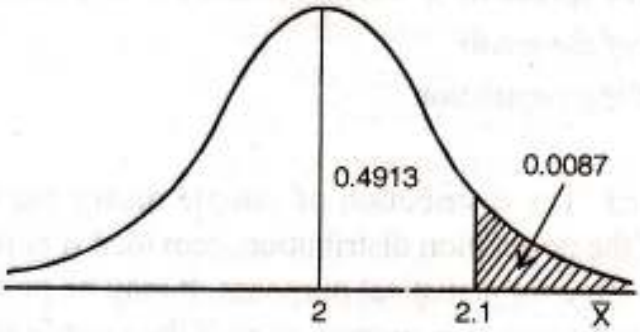
Solution. Since the population is normally distributed, therefore, the sampling distribution of the sample mean will al follow a normal distribution with mean

$$\mu_{\bar{x}} = \mu = 2$$

and standard error (s.d. of the sampling distribution)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.25}{\sqrt{36}} = 0.042$$

Therefore, the probability that the sample mean will be greater than 2.1 minutes is given by $P_r\ [\bar{x} \geq 2.1]$



To get the values from the standard normal distribution, this normal variate $\bar{x}$ must be converted into a standard normal variate by the transformation

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Therefore, the above probability statement becomes

$$P_r\left[\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \geq \frac{2.1 - \mu}{\sigma_{\bar{x}}}\right]$$

or

$$P_r\left[z \geq \frac{2.1 - 2}{0.042}\right]$$

or

$$P_r\ [z \geq 2.38].$$

From the table, this value of $z = 2.38$ corresponds to the area of 0.4913 to the left of the value $z = 2.38$. To get the required probability, this area of 0.4913 must be subtracted from the total area, *i.e.*,

$$P_r\ [\bar{x} \geq 2.1] = 0.5 - 0.4913 = 0.0087$$

Therefore, in only 0.87% of all possible sample of size $n = 36$, the sample mean will be greater than 2.1 minutes.

Distribution of Sample Medians

If a universe is large and can be approximated closely by a normal distribution with a mean μ and a standard deviation σ, the medians of random samples of size n are distributed with a mean μ and a standard deviation $1.2533\ \sigma / \sqrt{n}$, and the distribution of sample medians is nearly normal if n is large.

The standard deviation of the distribution of sample medians is called the *standard error of the sample median* and is denoted by :

$$\sigma_{Med} = 1.2533\ \sigma / \sqrt{n}$$

It should be noted that while the expectation of the median is same as expectation of the mean, the standard error of the median is greater than the standard error of the mean by a multiplier of 1.2533.

Distribution of Sample Standard Deviations

In the analysis of random variables relevant to business problems, it is common that the standard deviation of the population is unknown. In such a case, σ must be estimated by the sample standard deviation. This distribution is defined by a theorem that states :

If a population is large and normally distributed with a standard deviation σ, the standard deviation of the population is unknown. In such a case, σ must be estimated by the sample standard deviation. This distribution is defined by a theorem that states :

If a population is large and normally distributed with a standard deviation σ, the standard deviations of random samples of size n (where n is large), are closely approximated by a normal distribution with a standard deviation $\sigma / \sqrt{2n}$.

The standard deviation of the distribution of standard deviations of samples drawn from a normal population is called the *standard error of the standard deviation* and is denoted by :

$$S = \sigma / \sqrt{2n}$$

where S = Standard error of the standard deviations.

Sampling Distribution of the Difference of the Two Means

Suppose we have two populations, the first of size N_1, with mean μ_1 and standard deviation σ_1, and the second of size N_2, with mean μ_2 and standard deviation σ_2. The comparison is made on the basis of two independent random samples, with one of size n_1 drawn from the first population and the other of size n_2 drawn from the second population. If $\bar{x}_1$ and $\bar{x}_2$ are the two sample means, we can evaluate the possible difference between μ_1 and μ_2, by the difference of the sample means $\bar{x}_1 - \bar{x}_2$. The problem is one of determining the properties of a sampling distribution of $\bar{x}_1 - \bar{x}_2$.

The important properties of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ are :

1. With simple random sampling from two independent populations, the mean of the sampling distribution of $\bar{x}_1 - \bar{x}_2$, denoted by $\mu_{\bar{x}_1 - \bar{x}_2}$ is equal to the difference between the population means, *i.e.*,

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

2. The standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ (also known as standard error of $\bar{x}_1 - \bar{x}_2$) is given by

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(since $\bar{x}_1$ and $\bar{x}_2$ are independent random variables, the variance of their difference is the sum of their variances).

3. If $\bar{x}_1$ and $\bar{x}_2$ are the means of two independent samples drawn from two large or infinite populations, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be normal if the samples are of sufficiently large size.

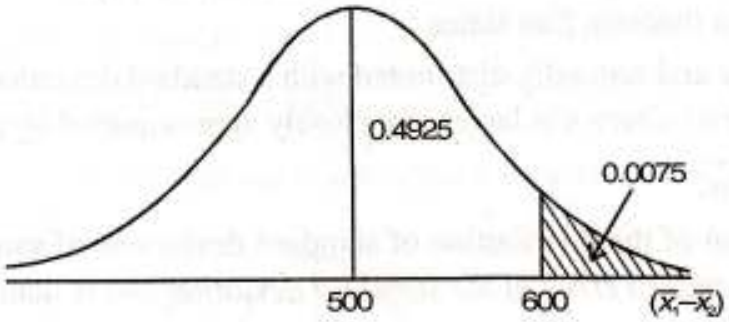
Illustration 2. Strength of the wire produced by company *A* has a mean of 4,500 kg and a standard deviation of 200 kg. Company *B* has a mean of 4,000 kg and a standard deviation of 300 kg. If 50 wires of company *A* and 100 wires of company *B* are selected at random and tested for strength, what is the probability that the sample mean strength of *A* will be (*i*) at least 500 kg more (*ii*) at least 400 kg more than that of company *B*. (*MBA, Delhi Univ., 2007*)

Solution. For the sampling distribution of the difference of two means, we have

$$\mu_{x_1 - x_2} = \mu_1 - \mu_2 = 4500 - 4000 = 500$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40000}{50} + \frac{90000}{100}} = \sqrt{800 + 900} = \sqrt{1700} = 41.23$$

The desired probability is given by $P_r(\bar{x}_1 - \bar{x}_2) \geq 600$ and is shown as shaded region below :



To convert this into standard normal variate, we get

$$P\left[\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \geq \left[\frac{600 - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right]\right.$$

or

$$P\left[Z \geq \frac{600 - 500}{41.23}\right] = P[Z \geq 2.43]$$

From the standard normal table, the corresponding value for Z = 2.43 is 0.4925. Hence, the required probability is given by

$$P(\bar{x}_1 - \bar{x}_2) \geq 600 = 0.5 - 0.4925 = 0.0075$$

Therefore, the probability that the sample mean strength of the wire produced by company *A* is greater than or equal to 600 kg than that of company *B* is given by 0.0075.

Sampling Distribution of the Number of Successes

If a random sample of size *n* is taken from a population whose elements belong to two mutually exclusive categories—one containing elements which possess a certain trait and the other containing elements which do not possess the trait—then the sampling distribution of the number of successes is the binomial distribution if sampling is made with replacement; and it is the hypergeometric distribution if sampling is made without replacement.

The sampling distribution of the number of successes being a binomial probability model will have its mean $\mu = np$ and standard error denoted by $\sigma = \sqrt{npq}$.

Illustration 3. If a coin is tossed 20 times and the coin falls on head after any toss, it is a success. Suppose the probability of success is 0.5. What is the probability that the number of successes is less than or equal to 12 ?

Solution : Given $\mu = np = 20 \times 0.5 = 10$, $\sigma = \sqrt{npq} = \sqrt{20 \times 0.5 \times 0.5} = \sqrt{5} = 2.24$

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{npq}} = \frac{12 - 10}{2.24} = 0.89$$

From the table, the corresponding value of $Z = 0.89$ is 0.8133. Hence, the probability that the number of successes is less than or equal to 12 is 0.8133.

Note : Since we are dealing with a proportion, the binomial distribution tends to normal distribution provided n is large enough to make both np and nq at least 5.

Sampling Distribution of Proportions

A population proportion is defined as $\pi = X/N$, where X is the number of elements which possess a certain trait and N is the total number of items in the population. A sample proportion is defined as $p = x/n$, where x is the number of items in the sample which possess a certain trait and n is the sample size. A proportion may be considered as a proportion of successes and is obtained by dividing the number of successes by sample size n. If a random sample of n is obtained with replacement, then the sampling distribution of p follows the binomial probability law.

Suppose that a population is infinite and that the probability of occurrence of an event (called its success) is π while the probability of non-occurrence of the event is $(1 - \pi)$. Now, consider all possible samples of size n drawn from this population and for each sample determine the proportion 'p' of successes. Then we obtain a *sampling distribution of sample proportion* whose mean is μ_p, and standard deviation σ_p are given by :

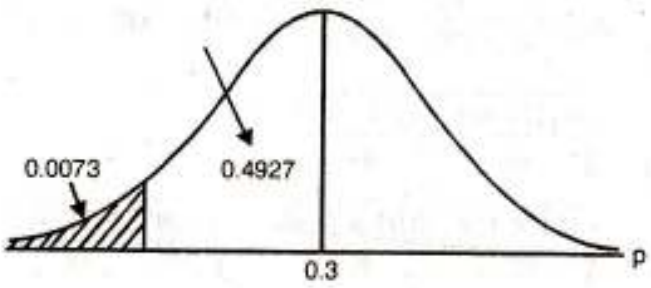
$$\mu_p = \pi \; ; \text{ and } \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

where σ_p = standard error of sample proportion. It measures chance variations of sample proportions from sample to sample. For large values of n $(n \geq 30)$ the sampling distribution is very closely approximated as normally distributed.

Illustration 4. In a quality department of manufacturing paints, at the time of despatch of decorative paints, 30% of the containers are found to be defective. If a random sample of 500 is drawn with replacement from the population, what is the probability that the sample proportion will be less than or equal to 25% defective ?

Solution. We have $\mu_p = \pi = 0.3$

and $$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.3 \times 0.7}{500}} = 0.0205$$



$$P[p \leq 0.25] = \left[Z \leq \frac{0.3 - 0.25}{0.0205}\right]; Z \leq 2.44$$

From the table, the corresponding value of Z is 0.4927, and therefore, the required probability is 0.0073 that sample proportion will be less than or equal to 0.25.

Illustration 5. In the year 2001, a policy is introduced to give loan to unemployed engineers to start their own business. Out of 1,00,000 unemployed engineers, 60,000 accept the policy and got the loan. A sample of 100 unemployed engineers is taken at the time of allotment of loan. What is the probability that sample proportion would have exceeded 50% acceptance ?

Solution. Here, $\mu_p = \pi = 0.60$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$= \sqrt{\frac{0.6 \times 0.4}{100}} \sqrt{\frac{1,00,000 - 100}{1,00,000 - 1}} = \sqrt{0.0024}\sqrt{0.999} = 0.0489$$

$$z = \frac{p-\pi}{\sigma_p} = \frac{0.50-0.60}{0.0489} = -2.04$$

The value of z can be found from the normal table and the corresponding value is 0.9793.

Therefore, the probability that sample proportion would have exceeded 50% acceptance is 0.9793. If all possible samples of 100 unemployed engineers are taken from the population of 1,00,000 then in 97.93% of these samples, the proportion of engineers who are in favour, is greater than 0.50.

Sampling Distribution of the Difference of Two Proportions

Earlier, in this chapter we referred to the sampling distribution of the difference of two means. Corresponding results can be obtained for the sampling distributions of difference of two proportions from two binomially distributed populations with parameters π_1 and π_2 respectively, when two random samples are drawn from two binomial populations and then compared. Unless both samples are of the same size, we cannot work with the number of successes, one must work only with the proportion of successes. Consider an example, a sample of 100 salesmen is taken from a chemical company, 50 are found to be in favour of new advertising policy. Another example of a textile company shows that 60 out of 150 are found to be in favour of policy. These two cases cannot be calculated unless they are reduced to proportions. The mean and standard deviation of this sampling distribution is given below :

$$\mu_{p_1-p_2} = \mu_{p_1} - \mu_{p_2} = \pi_1 - \pi_2$$

and

$$\sigma_{p_1-p_2} = \sqrt{\sigma^2_{p_1} + \sigma^2_{p_2}} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

If n_1 and n_2 are large ($n_1, n_2 \geq 30$), the sampling distributions of difference of two proportions are very closely normally distributed.

Illustration 6. Ten per cent of machines produced by company A are defective and five per cent of those produced by company B are defective. A random sample of 250 machines is taken from company A and a random sample of 300 machines from company B. What is the probability that the difference in sample proportion is less than or equal to 0.02 ?

(MBA, South Gujrat Univ.; MBA, Delhi Univ., 1999)

Solution. Under the assumption, the sampling distribution of $p_1 - p_2$ would have mean

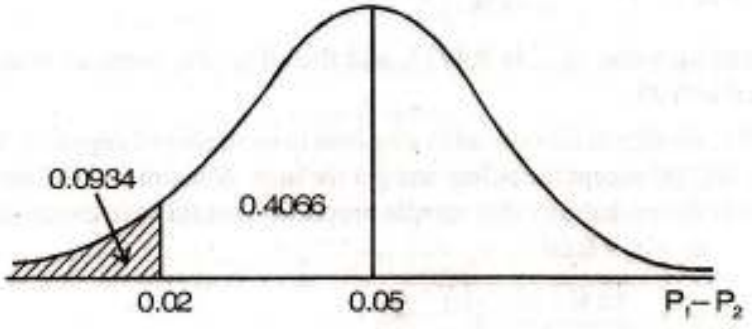
$$\mu_{p_1-p_2} = \pi_1 - \pi_2 = \frac{10}{100} - \frac{5}{100} = 0.1 - 0.05 = 0.05$$

and standard deviation

$$\sigma_{p_1-p_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

$$= \sqrt{\frac{0.1 \times 0.9}{250} + \frac{0.05 \times 0.95}{300}} = \sqrt{\frac{0.09}{250} + \frac{0.0475}{300}}$$

$$= \sqrt{0.00036 + 0.00016} = \sqrt{0.00052} = 0.0228$$



The probability that the difference in sample proportion is less than or equal to 0.02 is given by

$$P(p_1 - p_2) \leq 0.02$$

Hence, the required probability is obtained by transforming into a standard normal variate as

$$P\left[\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \le \frac{0.02-0.05}{0.0228}\right] \quad \text{or } P[z \le -1.32]$$

From the table, its corresponding value of z is 0.4066 and, therefore, the required probability is 0.0934 that the difference in sample proportion is less than or equal to 0.02.

We have discussed above few important types of sampling distributions. Just as we have discussed the sampling distributions of means, proportions, etc. It is also possible to discuss about the sampling distributions of first and third quartiles, quartile deviation, coefficient of skewness, coefficient of correlation, etc. However, they have not been discussed and only the formulae for standard error (S. E.) are given :

$$\text{S E . of quartiles or } \sigma_{Q_1} = \sigma_{Q_3} = \frac{1.3632\sigma}{\sqrt{n}}$$

$$\text{S.E. of Q.D. of } \sigma_{QD} = \frac{0.7867\sigma}{\sqrt{n}}$$

$$\text{S.E. of coefficient of skewness } \sigma_{sk} = \sqrt{\frac{3}{2n}}$$

$$\text{S.E. of coefficient of correlation } \sigma_r = \frac{1-r^2}{\sqrt{n}}$$

Sampling distributions occupy a place of great prominence in statistical theory. A sampling distribution shows that a statistic of a random sample may take on any set of values; but these values do not have the same probability of occurrence. Sampling distribution constitute the basis of testing hypothesis and enable us to evaluate the validity of statistical inferences.

MISCELLANEOUS ILLUSTRATIONS

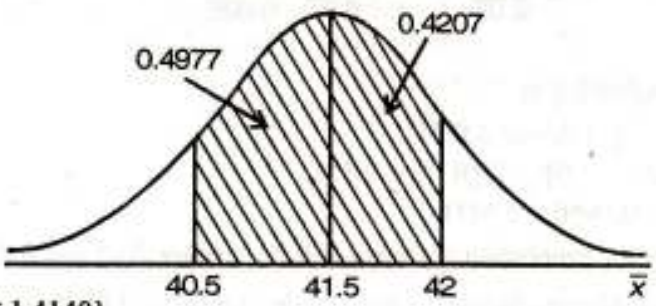
Illustration 7. The mean length of life of a certain cutting tool is 41.5 hours with a standard deviation of 2.5 hours. What is the probability that a simple random sample of size 50 drawn from this population will have a mean between 40.5 hours and 42 hours ? *(MBA, DU, 2005)*

Solution. Given, $\mu = 41.5, \sigma = 2.5, n = 50$

$$\sigma_{\bar{x}} = \frac{2.5}{\sqrt{50}} = \frac{2.5}{7.0711} = 0.3536$$

Therefore, the required probability is given by

$$P\{40.5 \le \bar{x} \le 42\} = P\left\{\frac{40.5-41.5}{0.3536} \le z \le \frac{42-41.5}{0.3536}\right\}$$



$= P\{-2.8281 \le z \le 1.4140\}$

$= P\{-2.8281 \le z \le 0\} + P\{0 \le z \le 1.4140\}$

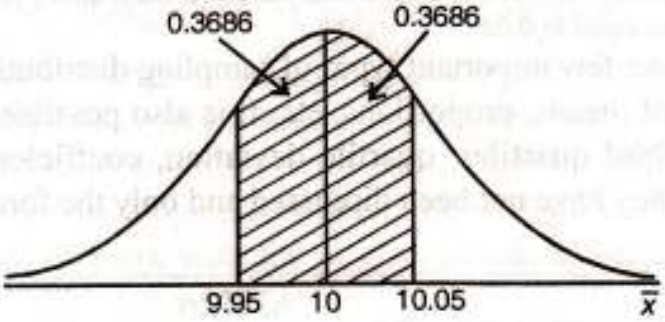
$= 0.4977 + 0.4207 = 0.9184.$

Therefore, the probability is 0.9184 for the cutting tool to have a mean life between 40.5 hours and 42 hours.

Illustration 8. A diameter of a component produced on a semi-automatic machine is known to be distributed normally with a mean of 10 mm and a standard deviation of 0.1 mm. If we pick up a random sample of size 5, what is the probability that the sample mean will be between 9.95 mm and 10.05 mm ?

Solution. Given : $\mu = 10$, $\sigma = 0.1$ and $n = 5$

$$z_1 = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{10.05 - 10}{0.1 / \sqrt{5}} = \frac{0.05}{0.0447} = 1.12$$



$$z_2 = \frac{9.95 - 10}{0.1 / \sqrt{5}} = -\frac{0.05}{0.0447} - 1.12$$

The area corresponding to Z = 1.12 is 0.3686. Hence, the required probability is 0.3686 + 0.3686 = 0.7372.

Illustration 9. A manufacturer of watches has determined from experience that 3% of the watches he produces are defective. If a random sample of 300 watches is examined, what is the probability that the proportion defective is between 0.02 and 0.035 ?

(MBA, Delhi Univ., 2000)

Solution. Here $\pi = 0.03$

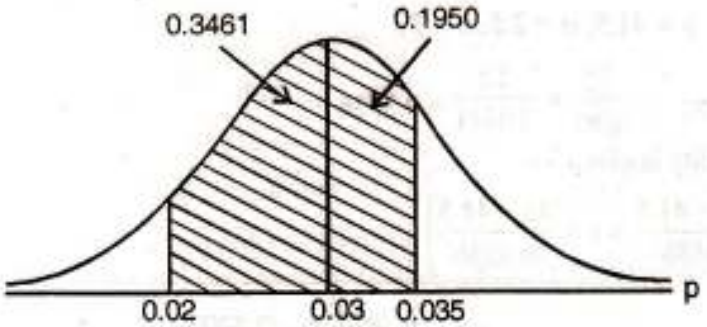
$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.03 \times 0.97}{300}} = \sqrt{\frac{0.0291}{300}} = \sqrt{0.000097} = 0.0098$$

When $p = 0.02$, then

$$z_1 = \frac{0.02 - 0.03}{0.0098} = \frac{-0.01}{0.0098} = 1.02$$

When $p = 0.035$, then

$$z_1 = \frac{0.035 - 0.03}{0.0098} = \frac{0.005}{0.0098} = 0.51$$



Therefore, the required probability is

$$P\{-1.02 \le z \le 0.51\}$$
$$= P\{-1.02 \le z \le 0\} + P\{0 \le z \le 0.51\}$$
$$= 0.3461 + 0.1950 = 0.5411$$

Hence, the probability that the proportion defective will be between 0.02 and 0.035 is given by 0.5411.

Illustration 10. Car Stereo of manufacturer *A* have mean lifetime of 1400 hours with a standard deviation of 200 hours, while those of manufacturer *B* have a mean lifetime of 1200 hours with a standard deviation of 100 hours. If random sample of 125 stereos of each manufacturer are tested, what is the probability that the manufacturer *A* stereos will have a mean lifetime which is at least (*i*) 160 hours more than the, manufacturer *B* stereos and (*ii*) 250 hours more than the manufacturer *B* stereos ?

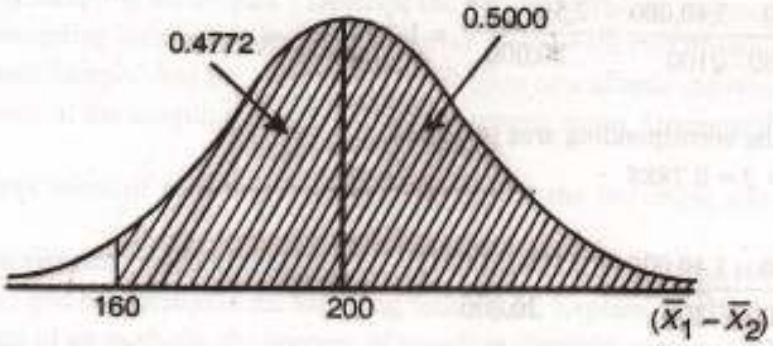
(MBA, Delhi Univ., 1999)

Solution. Let $\bar{x}_1$ and $\bar{x}_2$ denote the mean lifetime of samples A and B respectively. Then

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 = 1400 - 1200 = 200$$

and

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{125} + \frac{(100)^2}{125}} = \sqrt{320 + 80} = \sqrt{400} = 20$$



(i) The required probability is given by

$$= P\,(\bar{x}_1 - \bar{x}_2) \geq 160$$

To convert this into a standard normal variate, we get

$$P\left[\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \geq \frac{160 - 200}{20}\right] = P\,[z \geq -2]$$

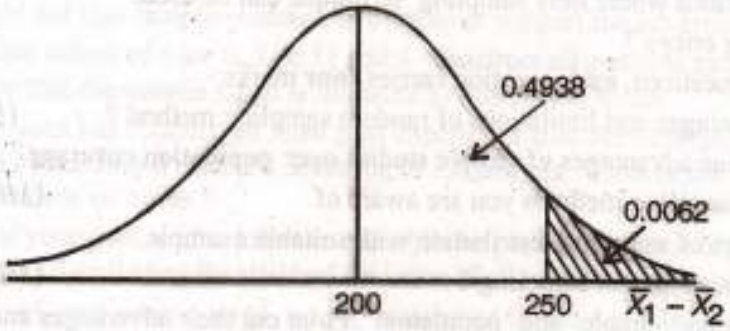
Required probability = area under normal curve to right of $z = -2$

$$= 0.5000 + 0.4772 = 0.9772$$

(ii) The required probability is given by

$$P\,[\bar{x}_1 - \bar{x}_2) \geq 250]$$

$$P\left[\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \geq \frac{250 - 200}{20}\right] = P\,[z \geq +2.5]$$



Required probability = area under normal curve to right of $z = 2.5$.

$$= 0.5000 - 0.4938 = 0.0062.$$

Illustration 11. The average annual starting salary for MBA (Marketing majors) is Rs. 3,40,000. Assume that for the population of MBA (Marketing majors), the average annual starting salary is $\mu = 3,40,000$ and the standard deviation is $\sigma = 20,000$. What is the probability that a simple random sample of MBA (Marketing majors) will have a sample mean within ± Rs. 2,500 of the population mean for each of the sample sizes : 50, 100 and 200 ? What is your conclusion ? *(MBA, Delhi Univ., 2003)*

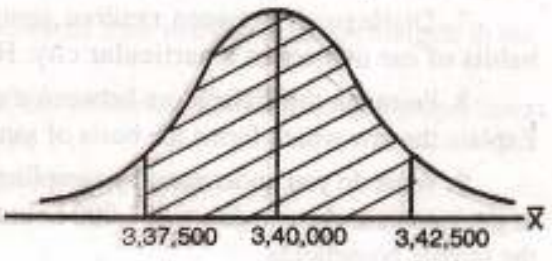
Solution : Given $\mu = 3,40,000$, $\sigma = 20,000$, $n_1 = 50$, $n_2 = 100$, $n_3 = 200$

For $n_1 = 50$

$$Z_1 = \frac{3,42,500 - 3,40,000}{20,000/\sqrt{50}}$$

$$= \frac{2,500 \times 7.071067}{20,000}$$

$$= \frac{17,677.67}{20,000} = 0.88$$



$$Z_2 = \frac{3,37,500 - 3,40,000}{20,000/\sqrt{50}} = -0.88$$

For $Z = \pm 0.88$, the corresponding area is 0.3106.

Required prob. $= 0.3106 \times 2 = 0.6212$

For $n_2 = 100$

$$Z_1 = \frac{3,42,500 - 3,40,000}{20,000/\sqrt{100}} = \frac{2,500 \times 10}{20,000} = 1.25$$

$Z_2 = -1.25$

For $Z = \pm 1.25$, the corresponding area is 0.3944.

Required prob. $= 0.3944 \times 2 = 0.7888$

For $n_3 = 200$

$$Z = \frac{3,42,500 - 3,40,000}{20,000/\sqrt{200}} = \frac{2,500 \times 14.142}{20,000} = 1.7676$$

$Z = -1.7676$

For $Z = \pm 1.7676$ corresponding area is 0.4616.

Required prob. $= 0.4616 \times 2 = 0.9232$.

PROBLEMS

1-A : Answer the following questions, each question carries **one** mark:

(*i*) What are the advantages of sampling ?

(*ii*) Define 'stratified random sampling'.

(*iii*) Explain probability sampling.

(*iv*) What is systematic random sampling ? *(M.A. Eco., M.K. Univ., 2003)*

(*v*) What is random sampling ?

(*vi*) How a stratified sample is selected ?

(*vii*) What is multi-stage sampling ?

(*viii*) What are the limitations of sampling ?

(*ix*) Mention few situations where only sampling technique can be used.

(*x*) What are sampling errors ?

1-B : Answer the following questions, each question carries **four** marks:

(*i*) What are the advantages and limitations of random sampling method ? *(M.A. Eco., Madras Univ., 2003)*

(*ii*) What are the various advantages of sample studies over population coverage.

(*iii*) Explain any four sampling methods you are aware of. *(MBA, Bharathidasan Univ., 2003)*

(*iv*) Explain the concept of sampling distribution with suitable example.

(*v*) What are the various types of sampling ? *(MBA, Bharathidasan Univ., 2001)*

(*vi*) Differentiate between 'Sample' and 'population'. Point out their advantages and limitations.

2. Point out the importance of sampling in solving business problems. What are the principles on which sampling theory rests ?

3. "Sampling is necessary under certain conditions". Explain this with illustrative examples.

4. Describe the various methods of sampling and the requisites of a good sample.

5. What is sampling ? Explain the importance of sampling in solving business problems. Critically examine the well known methods of probability sampling and non-probability sampling.

6. Define judgment sampling, quota sampling, and convenience sampling. Under what conditions, can each of these designs be used to advantage ?

7. Distinguish between random sampling and stratified sampling. Suppose it is desired to survey petrol buying habits of car owners in a particular city. How would you proceed about it ?

8. Point out the differences between a sample survey and a census survey. Under what conditions, are these undertaken ? Explain the law which forms the basis of sampling.

9. What do you understand by sampling ? In order to determine a new cost of living index, it is proposed to make a survey of the income and expenditure of 1,000 households in a large city. Describe carefully two methods which might be used to select the sample households.

10. Suppose you are asked to conduct a survey on the smoking habits of the Delhi University teachers. How will you proceed ?
11. "In any sample survey there are many sources of errors. A perfect survey is a myth." Discuss the statement.
12. "Data collected in census are automatically free of errors." Discuss the validity of the statement.
13. Enumerate the various methods of sampling and describe two of them mentioning the situations where each one is to be used.
14. What is the importance of sampling techniques ? Describe the various sampling techniques.
15. Explain the concepts of sampling distribution and standard error. Discuss the role of standard errors in large sample theory.
16. Explain the terms 'Random Sample' and the 'Sampling Distribution of a sample statistic'.
17. Find the mean and variance of the sampling distribution of the sample mean. Distinguish between standard deviation and standard error.
18. "There are many different ways of selecting a sample." Describe the important sampling methods pointing out the characteristics of each.
19. (*a*) Distinguish between sampling and non-sampling errors. What are their sources ? How these errors can be controlled ?
    (*b*) List the probabilistic and non-probabilistic sampling techniques. Explain stratified random sampling technique.
    (*c*) Explain with the help of an example, the concept of sampling distribution of a sample statistic and point out its role in managerial decision-making. *(MBA, Delhi Univ., 2003)*
20. The weight of certain type of a car tyre is normally distributed with a mean of 25 pounds and variance of 3 pounds. A random sample of 50 tyres is selected. What is the probability that the mean of this sample lies between 24.5 and 25.5 pounds ?
    [0.9586]
21. For a particular brand of T.V. picture tube, it is known that the mean operating life of the tubes is 1,000 hours with a standard deviation of 250 hours. What is the probability that the mean for a random sample of size 25 will be (*i*) greater than 1,000 hours, (*ii*) less than 1,000 hours, (*iii*) between 950 and 1,050 hours ? *(MBA, DU, 2005, 2006)*
    [0.5, 0.5, 0.6826]
22. An auditor takes a sample of size 36 from a population of 1,000 accounts receivable. The standard deviation of the population is unknown, but the standard deviation of the sample is Rs. 100. If the true mean value of the accounts receivable from the population is Rs. 3000, what is the probability that the sample mean will be less than or equal to Rs. 2800 ?
23. A manufacturer of razor blades claims that his product will, on the average, give 15 good shaves. Suppose you have five friends who try using one of these razor blades each. The number of shaves reported by your friends are 12, 16, 8, 14 and 10.
    (*a*) Find the mean and standard deviation of this sample.
    (*b*) Suggest how you might use this sample evidence to dispute or support the advertiser's claim.
24. For a population of size 5, the values of x are 8, 3, 1, 11 and 4. Construct all possible sample of size two and calculate their sample means. Hence, show that the sample mean is the same as population mean.
25. A manufacturer of knitting yarn has established from past experience that the breaking strength of this yarn is normally distributed with a mean of 12 pounds and standard deviation of 1.8 pounds. What is the probability that a sample size of 49 yield a mean of 14.5 pounds or more ?
26. Design a simple example of your own to illustrate the use of finite population correction factor by listing your values of some population, finding σ, and then finding the standard deviation of all possible sample of size 3 drawn without replacement. Does the standard deviation of your sample equal $\sigma / \sqrt{n}$ multiplied by the population correction factor ?
27. Two methods of performing a certain task in a manufacturing plant, method A and method B, are under study. The variable of interest is length of time required to perform the task. It is known that the variance of method A is 9 minutes squared and variance of method B is 12 minutes squared. A simple random sample of 35 employees performed the task by method A and independent simple random sample of 35 employees performed the task by method B. The average time required by the first group to complete the task was 25 minutes and the average time for the second group was 23 minutes. What is the probability of observing difference this large, if there is no difference in the true average length of time required to perform the task ?
28. An accountant has determined from prior experience that 60 per cent of his client's customers respond to initial requests for confirmation of their account balances. If a simple random sample of 64 customers is sent requests for confirmation, what is the probability that 50 per cent or more will respond ?
29. A research group stated that 16 per cent of the firms of a particular type, A increased their marked research budgets in the five years preceding the study. For type B firms the figure was 9 per cent.
    (*a*) What are the mean and standard deviation of the sampling distribution of the difference between sample proportions based on independent simple random samples of 100 firms from each type ?
    (*b*) What proportion of the sample differences would be between 0.05 and 0.10 ?
    (*c*) If you took a simple random sample of size 100 from each industry, what is the probability that the difference you would observe would be equal to or less than 0.02 ?

30. Suppose it is known that 5 per cent of forms processed by a clerical pool contain at least one error. If a simple random sample of 475 forms is examined, what is the probability that the proportion containing at least one error will be between 0.03 and 0.075 ?
31. A manufacturer of pens has determined from experience that 4 per cent of the pens he produces are defective. If a random sample of 400 pens is examined, what is the probability that the proportion defective is between 0.025 and 0.048 ?
32. Marks obtained by a number of students are assumed to be normally distributed with mean 65 and variance 25. If 3 students are taken at random, what is the probability that exactly two of them will have marks over 70 ?
33. A firm produces light bulbs that are known to have a mean life time of 1,200 hours with a standard deviation of 210 hours. What is the probability that a simple random sample of 100 bulbs will yield a mean that falls between 1,140 and 1,260 hours ?
[0.9956]

*****

14

Estimation of Parameters

INTRODUCTION

One important problem of statistical inference is the estimation of population parameters (such as population mean, variance, etc.) from the corresponding sample statistics (*i.e.*, sample mean, variance, etc.). There are several occasions, on which, we have to estimate population values from sample data in order to make a business decision. For example, a firm may wish to estimate the average amount of time its salesmen spend on each sales call; the telephone department may be interested in estimating the average length of a conversation for a long distance telephone call; a company may be interested in estimating the share of the population aware of its products. If all these estimates are obtained on a census basis, it would be very costly and a time-consuming proposition. Hence, quite often, sampling theory is employed to obtain information about samples drawn at random from a known population and an attempt is made to infer information about a population by use of samples drawn from it.

Statistical Estimation is the procedure of using a sample statistic to estimate a population parameter. A statistic used to estimate a parameter is called an *estimator* and the value taken by the estimator is called an *estimate*. Statistical estimation is divided into two main categories: Point Estimation and Interval Estimation.

Point Estimation. An estimate of a population parameter given by a single number is called a *point estimate* of the parameter. For example, if a firm takes a sample of 50 salesmen and finds out that the average amount of time each salesman spends with his customers is 80 minutes. If this figure is used for an estimate for all the salesmen employed by the firm, it is referred to as a "Point estimate" because we are using one value to obtain the population value. If one must rely on a single value as an estimate of a parameter, it is desirable to select the random variable that is expected to provide the most dependable estimate. Now, suppose there are several alternative estimators which might be used for estimating the same parameter. For example, the population parameter may be a measure of central tendency of the population values; then sample mean, sample median and sample mode may be considered as the possible estimators of the population parameter. Which sample statistics should be used as the estimator of the population parameter ? We need to establish "criteria" for choosing a satisfactory estimator which tell us which statistic does the "best" job of estimating it. The best estimator is the one that is more suitable to a given problem, most likely to give the desired result, is the least risky, and has such desirable properties as of unbiasedness, consistency, efficiency and sufficiency which are discussed in detail in this chapter.

Properties of a Good Estimator

A good estimator, is one which is "close" to the population parameter being estimated. Some of the desirable properties of an estimator are :

(1) Unbiasedness,

(2) Consistency,

(3) Efficiency, and
(4) Sufficiency.

Unbiasedness. An estimator is a random variable as it is always a function of sample values. Then, if the average of these sample values is equal to the population parameter, then, it is unbiased estimate. Thus, an estimator is said to be unbiased if the expected value of the estimator is equal to the population parameter being estimated. If θ is the parameter being estimated and $\hat{\theta}$ (read "theta hat") is an unbiased estimator of θ, then

$$E(\hat{\theta}) = \theta$$

For example, the sample mean is an unbiased estimator of the population mean, since

$$E(\bar{x}) = E\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right]$$

$$= \frac{1}{n}\left[E(x_1) + E(x_2) + \ldots + E(x_n)\right]$$

$$= \frac{1}{n} n\mu = \mu$$

Similarly, consider a problem of estimating p, the population proportion of successes in a binomial distribution. It can be shown that, if a sample yields x successes is n trials, then the ratio x/n is an unbiased estimate of p. Since

$$E\left[\frac{x}{n}\right] = \frac{1}{n}E[x] = \frac{1}{n}np = p$$

Therefore, the sample proportion is an unbiased estimator of population proportion.

If the sampling distribution of $\hat{\theta}$ is such that

$$E(\hat{\theta}) \neq \theta$$

Then the estimator is said to be *biased*. The bias is the quantity by which $E(\hat{\theta})$ and θ differ. The sample variance s^2 computed with the division constant $1/n$ is a biased estimator of σ^2, because

$$E(s^2) = \left(1 - \frac{1}{n}\right)\sigma^2$$

Here, the bias is the quantity $-\frac{\sigma^2}{n}$.

Hence, sample variance is not an unbiased estimator of population variance. But, if sample variance is computed with a division constant $1/(n-1)$, then it can be shown that $E(s)^2 = \sigma^2$, and therefore, s^2 is an unbiased estimator of σ^2, howsoever small may be the sample size. It may be pointed out that although s^2 is an unbiased estimator of σ^2, s is not an unbiased estimator of σ. The bias, however, diminishes rapidly as n increases.

Consistency. As the sample size increases, the difference between the sample statistic and the population parameter should become smaller and smaller. If the difference continues to become smaller and smaller as the sample size becomes larger, the sample statistic is said to converge in probability to a parameter is said to be *consistent* estimator of that parameter. Symbolically, if $\hat{\theta}$ is a sample statistic computed from a sample of size n and θ is the parameter being estimated. If

$$Pr\left[\left|\hat{\theta} - \theta\right| \leq d\right] \to 1 \qquad \text{as } n \to \infty$$

for any positive arbitrary d, then $\hat{\theta}$ is said to be a consistent estimator of θ.

This is true for $\bar{x}$ and s^2 which are consistent estimators of μ and σ^2 respectively. The sample median is a consistent estimator of the population mean only if the population distribution is symmetrical.

Efficiency. If the variance of the estimator is small, the distribution of the estimator will be better in that its value will be closer to the parameter value. This is the notion of efficiency. Efficiency can be treated as a relative term. In a sense all estimators are efficient; however, some estimators are more efficient than others. The efficiency of an estimator depends on its variance. A measure of relative efficiency can be computed by taking the ratio of the variances of two estimators of interest. If $\hat{\theta}_1$ is an unbiased estimator of $\hat{\theta}$, and $\hat{\theta}_2$ represents another unbiased estimator of θ, then the relative efficiency of $\hat{\theta}_1$ to $\hat{\theta}_2$ is given by

$$\text{Relative efficiency} = \frac{\text{Var}[\hat{\theta}_2]}{\text{Var}[\hat{\theta}_1]}$$

For a symmetrical distribution, both the sample mean and sample median are unbiased and consistent estimators of the population mean. We choose between them on the basis of relative efficiency, we select that one which has the smaller variance. It can be proved that sample mean $\bar{x}$ is preferred to the sample median as an estimator of μ because $\bar{x}$ is more efficient estimator.

Sufficiency. The fourth and last property of a good estimator that was developed by a famous statistician, Sir R.A. Fisher, is sufficiency.

A sufficient estimator is one that uses all information about the population parameter contained in the sample. For example, the sample mean, $\bar{x}$, is a sufficient estimator of the population mean since all the information in the sample is used in its computation. On the other hand, sample mid-range is not a sufficient estimator since it is computed by averaging only the highest and lowest values in the sample. A sufficient estimator ensures that all information that a sample can furnish with respect to the estimation of a parameter is being utilized. It may be noted that $\bar{x}$, p, $(\bar{x}_1 - \bar{x}_2)$ and $(p_1 - p_2)$ are sufficient estimators of the corresponding parameters μ, π, $(\mu_1 - \mu_2)$ and $(\pi_1 - \pi_2)$ respectively. A primary importance of the property of sufficiency is that it is a necessary condition for efficiency.

It is desirable that an estimator has all the properties discussed above. However, in practice, it is not always possible to determine such an estimator. It may be noted that though the point estimates are easy to use and understand, they cannot tell us how close we are to the true population value when used by themselves. Even the best point estimate may deviate enough from the parameter value to make the estimate unsatisfactory.

We have discussed above the properties desirable of an estimator to possess. We shall now discuss one of the most important methods called method of maximum likelihood that may provide estimators satisfying these properties.

Method of Maximum Likelihood

A general method for determining good estimators is called the *Method of Maximum Likelihood*. If a parameter θ is viewed as a variable, the method of maximum likelihood leads to the selection of a value of θ such that the likelihood (probability) of randomly obtaining a set of sample values is a maximum.

If $\bar{x}$ is a discrete random variable having a probability function $f(x\,;\theta)$ with only one parameter, θ, the likelihood function of the random sample $x_1, x_2, \ldots\ldots, x_n$ is

$$L(x_1, x_2, \ldots\ldots, x_n/\theta) = f(x_1\,;\theta)\, f(x_2\,;\theta) \ldots\ldots f(x_n\,;\theta).$$

The general rule followed in finding that value which maximizes the likelihood function is to

differentiate the likelihood function with respect to the parameter θ, set the resultant function equal to zero, and solve. Hence, the value of the parameter θ that maximizes the likelihood function is the maximum likelihood estimate.

The logic of this method can more easily be grasped by an example. Random sampling is based on the assumption that a sample, as a small scale replica of the population, will tend to reflect the properties of the population. For example, if a random sample of 35 students is selected from a class of 100 and average weight is calculated, we might get average weight as 120 lbs. One might ask what is the most likely value of μ in view of the sample result ? The conclusion is that the most likely value of μ is $\bar{x}$ = 120 lbs. and not 140 or 110 lbs. or some other value. The value of μ is indeed a definite, fixed value; but the method of maximum likelihood views μ, as if, it were a variable such that the most likely value of μ is, its *maximum likelihood estimator.*

The method of maximum likelihood provides estimators which are consistent, efficient and sufficient but it does not always provide estimators that are unbiased.

Illustration 1. A market organisation wants to introduce lottery systems to promote sales. The manager surveyed a sample of ten consumers for introduction of this system. Out of ten, six are in favour of this system. What is reasonable estimate of the population proportion π ?

Solution : Sample proportion $p = \frac{x}{n} = \frac{6}{10} = 0.6$. We would like to know, given, $p = 0.6$, how likely is it that true proportion π = 0.1, 0.2, ... 0.9.

By the binomial distribution, the probability of x successes is given by π is

$$B(x ; \pi) = {}^{n}C_{x}\, \pi^{x} (1-\pi)^{n-x}$$

Suppose, the true population proportion is π = 0.3, this probability of obtaining sample with 6 successes in 10 is

$$B(6 ; 0.3) = {}^{10}C_{6}\, (0.3)^{6} (0.7)^{4} = 0.0368$$

That is, if π = 0.3, the chances are less than 4 in a hundred that this sample result would occur.

If π = 0.5, we obtain a likelihood of

$$B(6 ; 0.5) = {}^{10}C_{6}\, (0.5)^{6} (0.5)^{4} = 0.2051$$

Similarly, we can consider all other possible values of the parameter to determine how likely it is that the parameter considered would yield the sample actually observed.

We find that for π = 6.0, it has maximum probability 0.2508 which is the maximum likelihood estimate.

Interval Estimation. An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called as *interval estimate* of the parameter. Interval estimates indicate the precision or accuracy of an estimate and are, therefore, preferable to point estimates. The interval estimate or a "confidence interval" consists of an upper confidence limit and lower confidence limit, and we assign a probability that this interval contains the true population value. The first step in constructing a confidence interval is to decide how much confidence we want that this interval will contain the population value. Let us say, that we want 95 per cent confidence. This is known as a "95 per cent confidence level."

The previous chapter showed that when sampling is from a normally distributed population, the sampling distribution of sample mean is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. Knowing that $\bar{x}$ is normally distributed allows us to make additional statements about the distribution of $\bar{x}$. Thus, the sampling distribution of $\bar{x}$ can be transformed into the standard normal distribution by

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

Suppose we take the point $\mu \pm 1.96\ \sigma/\sqrt{n}$

Let $\bar{x} = \mu + 1.96\ \sigma/\sqrt{n}$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1.96\sigma/\sqrt{n}}{\sigma/\sqrt{n}} = 1.96$$

So, we can actually view z as the number of standard deviations a point is from μ. Since a z value of 1.96 corresponds to the point $\mu + 1.96\ \sigma/\sqrt{n}$, and z of -1.96 corresponds to $\mu - 1.96\ \sigma/\sqrt{n}$, clearly the area of the sampling distribution of $\bar{x}$ in the interval

$$\mu - 1.96\ \sigma/\sqrt{n} \text{ to } \mu + 1.96\ \sigma/\sqrt{n}$$

is 0.95. Therefore, 95 per cent of all sample means $\bar{x}$ are contained in this interval. Thus, the probability of drawing a random sample of size n and obtaining an $\bar{x}$ in this interval is 0.95.

$$P[\mu - 1.96\ \sigma/\sqrt{n} \le \bar{x} \le \mu + 1.96\ \sigma/\sqrt{n}] = 0.95$$

Similarly, it can be shown that

$$P[\mu - 2.578\ \sigma/\sqrt{n} \le \bar{x} \le \mu + 2.578\ \sigma/\sqrt{n}] = 0.99$$

The numbers 1.96, 2.578, etc., in the confidence limits are called *confidence coefficients* or *critical values*. It may be noted that the true mean may be expected to be no farther away than $3\sigma_{\bar{x}}$ from the sample mean that is a range of $\bar{x} \pm 3\sigma_{\bar{x}}$ will include the unknown true mean. Thus, the procedure in interval estimate comprises 3 steps :

1. The particular statistic, say, the mean of the sample or standard deviation of the sample is determined.
2. The confidence level is decided, *i.e.*, 95%, 99%, etc.
3. The standard error of the particular statistic is calculated.

Finally, we state, with a known degree of confidence, that the parameter is included in this interval.

Confidence Limits for Population Mean

Confidence limits for estimation of population mean μ are given by sample statistic $\pm z_c$ (S.E.) where z_c is critical value of z. 95% confidence limit for estimation of the population mean μ are given by $\bar{x} \pm 1.96\sigma_{\bar{x}}$; where lower confidence limit = $L = \bar{x} - 1.96\sigma_{\bar{x}}$ and upper confidence limit = $U = \bar{x} + 1.96\sigma_{\bar{x}}$.

Similarly, 99% confidence limits will be given by $\bar{x} \pm 2.58\sigma_{\bar{x}}$. For all practical purposes, a range of plus and minus three standard errors attached to sample mean, that is $\bar{x} \pm 3\sigma_{\bar{x}}$, will include the unknown true mean. More generally, the confidence limits are given by $\bar{x} \pm z_c\sigma_{\bar{x}}$ where z_c depends on the particular level of confidence desired. However, this is true in case sampling is from an infinite population or, if sampling is with replacement from a population of finite size N then confidence limits are given by :

$$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Illustration 2. The quality department of a wire manufacturing company periodically selects a sample of wire specimens in order to test for breaking strength. Past experience has shown that the breaking strengths of a certain type of wire are normally distributed with standard deviation of 200 kg. A random sample of 64 specimens gave a mean of 6,200 kg. The quality control supervisor wanted a 95 per cent confidence interval for the mean breaking strength of the population.

Solution. The z_C value corresponding to a confidence coefficient of 0.95 is 1.96. Therefore, the limits are :

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 6{,}200 \pm 1.96 \frac{200}{\sqrt{64}}$$

$$= 6{,}200 \pm 1.96 \times 25 = 6151 \text{ to } 6249$$

Hence, the 95 per cent confidence limits are 6151 to 6249.

Illustration 3. A manager wants an estimate of average sales of salesman in his company. A random sample of 100 out of 500 salesmen is selected and average sales is found to be Rs. 750 (thousand). If population standard deviation is Rs. 150 (thousand), manager specifies a 98% level of confidence. What is the interval estimate for average sales of salesman ?

Solution. Here $N = 500$, $n = 100$, $\bar{x} = 750$ and $\sigma = 150$.

The confidence limits are given by

$$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

where $z_c = 2.33$, and $\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{150}{100} \sqrt{\frac{400}{499}} = 15\,(0.895) = 13.425$

The required confidence interval is

$$750 \pm 2.33(13.425)$$

Thus, it can be stated that for 0.98 level of confidence, the population mean falls within the interval Rs. 718720 to Rs. 781280.

Confidence Limits for Population Proportion

If sampling is from an infinite population, or if sampling is with replacement from a finite population, the confidence limits for the population proportion are given by :

$$p \pm z_c \sqrt{\frac{pq}{n}}$$

where p is the proportion of success in the sample of size n.

If we are interested in calculating the 95% confidence limits for the population proportion, they would be given by :

$$p \pm 1.96 \sqrt{\frac{pq}{n}}$$

Illustration 4. The Human Resource director of a large organisation wanted to know what proportion of all persons who had ever been interviewed for a job with his organisation had been hired. He was willing to settle for 95 per cent confidence interval. A random sample of 500 interview records revealed that 76 or 0.152 of the persons in the sample, had been hired.

Solution. The 95 per cent confidence interval for the population proportion is given by $p \pm 1.96 \sqrt{\frac{pq}{n}}$

$$= 0.152 \pm 1.96 \sqrt{\frac{0.152 \times 0.848}{500}} = 0.152 \pm 0.032 = 0.12 \text{ to } 0.184.$$

Hence, the required proportion varies between 0.121 and 0.183.

Illustration 5. Out of 20,000 customers' ledger accounts, a sample of 600 accounts was taken to test the accuracy of posting and balancing wherein 45 mistakes were found. Assign limits within which the number of defective cases can be expected at 5% level of confidence.

Solution. We are given that

$$n = 660,\ p = \text{proportion of mistakes} = \frac{45}{600} = 0.075$$

$$q = 1 - p = 1 - 0.075 = 0.925$$

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.075 \times 0.925}{600}} = 0.011$$

Therefore, 95% confidence limits are given by

$$p \pm z_c \sqrt{\frac{pq}{n}}$$

$$= 0.075 \pm 1.96\ (0.011) = 0.075 \pm 0.022 = 0.053 \text{ to } 0.097$$

Hence, it is expected that the number of mistakes would vary between 5.3% and 9.7% at 5% level of significance. Here, the number of defective cases in a lot of 20,000 are expected to be between 20,000 × 0.053 and 20,000 × 0.097 or 1060 and 1940.

Confidence Limits for Difference of Two Means

When two independent random samples of $n_1 > 30$ and $n_2 > 30$ are taken then the sampling distribution of the difference of the two sample means $\bar{x}_1 - \bar{x}_2$ is approximately normal with mean $(\mu_1 - \mu_2)$ and $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Given the confidence coefficient, the approximate confidence limits for $(\mu_1 - \mu_2)$ may be expressed as follows :

$$(\bar{x}_1 - \bar{x}_2) \pm z_c\ \sigma_{\bar{x}_1 - \bar{x}_2}$$

Illustration 6. A sample of 150 items from machine *A* had an average life of 1,400 hrs. A similar sample of 100 items from machine *B* had a mean life of 1,200 hrs. Past records indicate that the standard deviation of the items produced by machine *A* is 120 hrs. and by machine *B* is 80 hours. Find 95 per cent confidence limits on the difference in the average lifetimes of the populations of the items produced by the two machines.

Solution. The 95 per cent confidence limits are given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_c\ \sigma_{\bar{x}_1 - \bar{x}_2}$$

where $\bar{x}_1 - \bar{x}_2 = 1400 - 1200 = 200$, and $z_c = 1.96$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(120)^2}{150} + \frac{(80)^2}{100}} = 12.6$$

Therefore, the required 95 per cent confidence limits are :

$$200 \pm 1.96\ (12.6) = 175.304 \text{ to } 224.696$$

Hence, the 95 per cent confidence limits are 175.304 to 224.696 for the difference in the average lifetime of the items produced by the two machines *A* and *B*.

Confidence Limits for Difference of Two Proportions

The confidence limits for the difference of two population proportions, where the populations are infinite, are given by :

$$(p_1 - p_2) \pm z_c \sigma_{p_1 - p_2}$$

where $(p_1 - p_2)$ = difference of proportions, and

$\sigma_{p_1 - p_2}$ = Standard error of the difference of two proportions given by :

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

We have discussed above the procedure of estimating a population measure for sample measure. The discussion has been confined to estimation for mean, proportion, difference of means, difference of proportions. These few parameters have been singled out for discussion because they are by far the most important decision parameters in so far as univariate data are concerned. However, it may be noted that the same general procedure of interval estimation can be applied to other parameters provided n is large enough for the central limit theorem to operate.

It may be noted that the procedure of estimation discussed in this chapter is applicable only in case of large samples (sample size greater than 30). Small samples require special treatment.

Determination of a Proper Sample Size

Thus far we have calculated the confidence intervals based on the assumption that the sample size n is known. In most of the practical situations, generally, it is not known. Instead, one may prefer to specify the width of the interval and use this information to solve for n. The method of determining a proper sample size, n is given for the following two cases :

(*a*) Sample size for estimating a population mean

The confidence interval formula is given by

$$\bar{x} \pm z_c \sigma / \sqrt{n}$$

or $$\bar{x} \pm E$$

where $$E = z_c \sigma / \sqrt{n}$$

is the maximum allowable sampling error, *i.e.*, difference between the population mean and the sample mean.

or $$\sqrt{n} = \frac{z_c \sigma}{E}$$

or $$n = \frac{z_c^2 \sigma^2}{E^2}$$

Here, both the values of z_c and E must be specified by the researcher, the value of the population σ may be actual or estimated.

Illustration 7. A cigarette manufacturer wishes to use a random sample to estimate the average nicotin content. The sampling error should not be more than one milligram above or below the true mean, with a 99 per cent confidence coefficient. The population standard deviation is 4 milligrams. What sample size should the company use in order to satisfy these requirements?

Solution. Here $E = 1$, $z_c = 2.58$, and $\sigma = 4$

Sample size formula is $$n = \frac{z_c^2 \sigma^2}{E^2}$$

Substituting the values, we get $$n = \frac{(2.58)^2 (4)^2}{1^2} = 106.50 \text{ or } 107$$

Hence, the required sample size is $n = 107$ which the company should use for their requirement to be fulfilled.

(*b*) Sample size for estimating a population proportion

The confidence interval formula for proportion is given by

$$p \pm z_c \sqrt{\frac{pq}{n}}$$

Using E to represent the maximum allowable sampling error, we may write the above equation as

$$p \pm E$$

where E is the difference between the sample proportion and the population proportion.

Now $$E = z_c \sqrt{\frac{pq}{n}}$$

Solving for n, we get

$$n = \frac{z_c^2 pq}{E^2}$$

where the values of z_c and E are predetermined. The value of population proportion p may be actual or estimated from the past experience.

Illustration 8. A firm wishes to estimate with a maximum allowable error of 0.05 and a 95 per cent level of confidence, the proportion of consumers who prefer its product. How large a sample will be required in order to make such an estimate if the preliminary sales reports indicate that 25 per cent of all consumers prefer the firm's product ?

Solution. Here $E = 0.05$, $p = 0.25$, and $z_c = 2.33$.

Substituting these values in the formula $n = \frac{z_c^2}{E^2} pq$, we get

$$n = \frac{(2.33)^2}{(0.05)^2}(0.25)(0.75)$$

$$= \frac{5.4289}{0.0025}(0.1875) = \frac{1.0179}{0.0025} = 407.16 \text{ or } 407$$

Hence the required sample size $n = 407$.

MISCELLANEOUS ILLUSTRATIONS

Illustration 9. A machine is producing ball bearings with diameter of 0.5 inches. It is known that the standard deviation of the ball bearings is 0.005 inches. A sample of 100 ball bearings is selected and their average diameter is found to be 0.498 inches. Determine the 99 per cent confidence interval.

Solution. Using the formula for confidence interval

$$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}}$$

where $\bar{x} = 0.498$, $z_c = 2.58$, $\sigma = 0.005$ and $n = 100$

We get $0.498 \pm 2.58 \frac{0.005}{\sqrt{100}} = 0.498 \pm 2.58\ (0.0005)$

$= 0.498 \pm 0.00129 = 0.4967$ to 0.4993

Hence, the 99 per cent confidence interval is 0.4967 to 0.4993 inches.

Illustration 10. Suppose we want to estimate the proportion of families in a town which has two or more children. A random sample of 144 families shows that 48 families have two or more children. Construct a 95 per cent confidence interval. *(MBA, HPU, 2007)*

Solution. Sample proportion $p = \frac{x}{n} = \frac{48}{144} = \frac{1}{3} = 0.333$

The confidence interval formula for proportion p is given by

$$p \pm z_c \sqrt{\frac{pq}{n}}$$

where $p = \frac{1}{3}$, $q = \frac{2}{3}$, $z_c = 1.96$, $n = 144$

and $\sqrt{\frac{pq}{n}} = \sqrt{\frac{(1/3)(2/3)}{144}} = 0.0393$

Therefore, the required confidence interval is

$0.333 \pm 1.96\ (0.0393) = 0.333 \pm 0.077 = 0.256$ to 0.410.

Hence, the population proportion of families who have two or more children is between 25.6 and 41.0 per cent.

Illustration 11. A ball pen manufacturer makes a lot of 10,000 refills. The procedure desires some control over these lots so that no lot will contain an excess number of defective refills. He decides to take a random sample of 400 refills for inspection from a lot of 10,000 and finds 9 defectives. Obtain a 90% confidence interval for the number of defectives in the entire lot.

Solution. The formula to be used for confidence interval is

$$p \pm z_c \sqrt{\frac{pq}{n}}$$

where $p = \frac{x}{n} = \frac{9}{400} = 0.0225$, $z_c = 1.645$

and $\sqrt{\frac{pq}{n}} = \sqrt{\frac{0.0225 \times 0.9775}{400}} = 0.0074$

Therefore, the required confidence interval is given by

$$0.0225 \pm 1.645(0.0074) = 0.0225 \pm 0.0122 = 0.0103 \text{ to } 0.0347$$

Hence, we may conclude with 90 per cent confidence that the population contains between 1.03 and 3.47 per cent defective in the entire lot.

Illustration 12. In a large consignment of oranges, a random sample of 500 oranges revealed that 65 oranges were bad. Prove that 99.73% of bad oranges in the consignment certainly lie between 8.5% and 17.5%.

Solution. Given that $n = 500$

p = number of bad oranges in the consingment $\frac{65}{500} = 0.13$, $q = 1 - p = 1 - 0.13 = 0.87$, $z_c = 3$

and $$\sqrt{\frac{pq}{n}} = \sqrt{\frac{0.13 \times 0.87}{500}} = 0.015$$

The 99.73% confidence limits for the population proportion of bad oranges in the consignment are given by

$$p \pm 3\sqrt{\frac{pq}{n}} = 0.13 \pm 3 \times 0.015 = 0.13 \pm 0.045 = 0.085 \text{ and } 0.175$$

Hence, the percentage of bad oranges in the consignment certainly lies between 8.5% and 17.5%.

Illustration 13. 400 labourers were selected at random from a certain city. Their mean income was Rs. 1700 per month with a standard deviation of Rs. 140. Set up 95% confidence limits within which the income of the labour community of the district is expected of lie.

Solution. Given $\bar{x} = 1700, \sigma = 140\ n = 400$ and $z_c = 1.96$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{140}{20} = 7$$

Therefore, 95% confidence limits are given by

$$\bar{x} \pm z_c \sigma_{\bar{x}} = 1700 \pm 1.96\ (7) = 1700 \pm 13.72$$
$$= 1686.28 \text{ to } 1713.72.$$

Illustration 14. In an attempt to control the quality of output for a manufactured part, a sample of parts is chosen randomly and examined in order to estimate the population proportion of parts that are defective. The manufacturing process operates continuously unless it must be stopped for inspection or adjustment. In the latest sample of 90 parts, 15 defectives are found. Determine the following estimates of π the population proportion defective (*a*) a point estimate (*b*) 98 per cent interval estimate.

Solution. (*a*) Point estimate : $p = \frac{15}{90} = 0.167$

(*b*) Interval estimate : $$\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.167)(0.833)}{90}} = 0.0393$$

98 per cent interval estimate shall be given by

$$p \pm z_c \sigma_p = 0.167 \pm 2.33 \times 0.0393$$
$$= 0.167 \pm 0.092 \text{ or } 0.075 \text{ to } 0.259.$$

Illustration 15. A random sample of 200 consumer accounts at a large brokerage firm is selected for the purpose of estimating the mean number of transactions per year for each customer. The sample mean is 12. Determine 99% confidence interval for the mean number of transactions of all consumer accounts of the firm.

Solution. Using the formula

$$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}}$$

where, $\bar{x} = 43, z_c = 2.58, \sigma = 2.5, n = 200$

$$43 \pm 2.58 \frac{2.5}{\sqrt{200}} = 43 \pm 0.456 = 42.544 \text{ to } 43.456.$$

PROBLEMS

1-A : Answer the following questions, each question carries **one** mark:

(*i*) What is statistical estimation ?

(*ii*) Distinguish between point estimate and interval estimate.

(*iii*) What is point estimation ? *(MA, Eco. M.K. Univ.)*

(*iv*) What are confidence limits for populaion mean ?

(*v*) What are confidence limits for population proportion ?

(*vi*) Name the important properties of a good estimator.

(*vii*) Give the formula for the method of maximum likelihood of a good estimator.

(*viii*) Which formula is used for determining confidence limits for difference of two means ?
(*ix*) Give the formula for determining sample size for estimating a population mean.
(*x*) What are confidence limits for difference of two proportions ?
(*xi*) Differentiate between confidence limits and confidence level.

1-B : Answer the following questions, each question carries **four** marks:

(*i*) Explain the concept of confidence interval with suitable example.
(*ii*) Briefly explain any two properties of a good estimator.
(*iii*) Describe the desirable properties of a good estimator.
(*iv*) What are confidence limits ? How are they determined ?
(*v*) How sample size is determined ? Explain with the help of an example.

2. What do you understand by estimation? In what sense, do we consider estimation as a procedure of decision-making ?
3. (*a*) What do you mean by 'Statistical Estimation'? Briefly explain the methodology used for estimating the mean of the population from the mean of the sample.
   (*b*) Distinguish clearly between the point estimation and interval estimation. In what way, do we say that an interval estimate is better than a point estimate ?
4. (*a*) Explain clearly the desirable properties of a point estimate.
   (*b*) What information and assumptions must be given to compute the sample size for an interval estimate of the universe mean ?
5. What is the difference between 'Statistic' and 'Parameter'? Explain, with examples, the methods employed for the estimation of population parameters based on sample means, difference of two means, sample proportion and difference of the sample proportions.
6. What is meant by confidence interval of a population parameter ?
7. With the help of an example, explain the method of maximum likelihood and point out its significance.
8. Comment on the statement, "Theoretically speaking, it is possible to have an estimate which is identical with the parameter being estimated. In practice, however, such an estimate is often unnecessary and physically impossible."
9. (*a*) Explain clearly the procedure involved in interval estimation.
   (*b*) Describe briefly the problems of estimation of population parameters.
10. Explain the following terms with the help of an example :
    (*i*) Confidence limits, (*ii*) Confidence interval,
    (*iii*) Interval estimate, (*iv*) Confidence coefficients or critical values.
11. What are the properties of a good estimator ? Prove that the mean of a simple random sample from a given population is an unbiased estimator of the population mean.
    (*a*) Explain briefly the properties of a good estimator.
    (*b*) Explain the concepts of (*i*) the power of statistical test, (*ii*) reliability and validity of measurements.
12. In a consignment of 1,00,000 tennis balls, 400 were drawn at random and examined. It was found that 20 of these were defective. How many defective balls can you expect in the whole consignment at 95% confidence level ?
13. A statistics consultant with the association of personnel director was asked to determine what proportion of electrical personnel who change jobs do so because they are bored with their work. A random sample of 400 electrical personnel who had recently changed jobs were enquired, and 200 stated that they changed jobs because they were bored. The statistician prepared a 95 per cent confidence interval for the true proportion changing jobs because of boredom. What are the lower and upper limits of this interval ?
14. A bank official is interested in knowing the difference between the average amount of money or deposit by customers in two branch banks. A random sample of 35 customers was selected from each branch. The sample means were as follows : Branch *A* : Rs. 4500; Branch *B* : Rs. 3250. The two populations are normally distributed with variances $\sigma_A^2 = 760$ and $\sigma_B^2 = 850$. Construct the 95 and 99 per cent confidence interval for $\mu_A - \mu_B$.
15. A random sample of 50 persons was interviewed to find their preference between two brands of tea. 35 of the interviewed persons preferred brand *A* to brand *B*. Find the 95 per cent confidence interval for the proportion of persons who prefer brand *A*.
16. After an intensive advertisement campaign of polish, the manufacturers wanted to know how many of the possible customers had read the advertisement. They selected a random sample of 50 customers and found that only 15 of them had read the advertisement. Find 95 per cent confidence interval for the proportion of customers who had not read the advertisement.
17. A manufacturer of television picture tubes tested 75 tubes to determine their mean lifetime. The sample yield an average of 4,200 hours with a standard deviation of 430 hours. Use a 95 per cent level of confidence for the interval estimate of the value below which the mean of the population should not fall.

18. A new drug has been developed for the treatment of a certain disease. A group of 400 patients suffering from the disease were treated with the new drug. Another group of 400 patients were treated with an alternative drug. At the end of two weeks, 320 of the patients receiving the new drug recovered, while 240 of those taking the alternative drug recovered. Construct the 95 per cent confidence interval for the difference in the true proportion of patients who might be expected to respond to the two drugs.
19. A sample of 16 observations has been taken from a population in which the random variable is normally distributed. The sample is 50 and the sample standard deviation is 10. Determine a 95 per cent confidence interval for the population mean.
20. A statistician is asked to conduct a survey to determine an estimate of the proportion of the people who favour the recall of the local politician. He is told that his estimate should not differ from the true proportion by more than 2 per cent with 95 per cent confidence. How large should his random sample be to produce an estimate of the proportion satisfying this condition?
21. The wearing quality of a certain type of truck tyre is to be estimated by road testing a sample of the tyres. It is estimated that the standard deviation of wearing quality is 200 km.
    (*a*) If the maximum allowable sampling error is 600 km, at a 95 per cent level of confidence, what should be the sample size ?
    (*b*) If the level of confidence were 99 per cent, what would be the appropriate sample size?
    (*c*) If the maximum allowable error were 300 km, what would be the appropriate sample size for a 95 per cent level of confidence ?
22. In measuring reaction time, a psychologist estimates that the standard deviation is 0.05 seconds. How large a sample of measurements must be taken in order to be 95% confident that the error of his estimate will not exceed 0.01 seconds ?
    [96]
23. A factory is producing 50,000 pairs of shoes daily. From a sample of 500 pairs, 20% were found to be of substandard quality. Estimate the number of pairs that can be reasonably expected to be spoiled in the daily production and assign limits at 5% level of significance.
    [Between 385 and 1,615]
24. The guaranteed average life of a certain type of electric light bulbs is 1,000 hours with a standard deviation of 125 hours. It is proposed to sample the output so as to ensure that 90% of the bulbs do not fall short of the guaranteed average by more than 2.5 per cent. What should be the minimum size of the sample ?
25. A random sample of six castings drawn from a universe of 75 castings shows the following weight for each. Compute an interval estimate for μ at 2% level of confidence.

Casting No. :	1	2	3	4	5	6
Weight (kg) :	82.9	83.5	84.1	83.6	82.5	84.4

26. 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 1/16, find 95% confidence limits of the percentage of defective items in the consignment.
27. From previous studies, the population standard deviation for a placement test has been determined to be 12.4. The test is scored on a scale of 0 – 100. A placement agency wants to be 90% confident that the average test score of a sample falls within plus or minus 3 points of the population average score. How large a sample should be selected ?
28. The foreman of a mining company has estimated the average quantity of ore extracted per shift to be 34.6 tons and the sample standard deviation to be 2.8 tons per shift based upon a random sample of six shifts. Construct 95% and 90% confidence around sample average estimate.
29. The life (in hours) of a 100 watt bulb is known to be normally distributed with standard deviation of 36 hours. A random sample of 15 bulbs yielded the following results :

    Life in hours

    2216 2237 2249 2204 2225 2301 2281 2263 2318 2255 2275 2295 2250 2238 2300

    Construct a 95% two sided confidence interval so that the actual mean of the life of bulbs fall within this interval.
30. A machine fills cans with a soft drink beverage and the manufacturer is interested in obtaining a confidence interval estimate of the variance of its fill volume. A random sample of 20 cans yields a sample variance of 0.0225. Construct a two-sided 95% confidence interval for variance.
31. A manufacturer is interested in estimating the proportion (p) of acceptable products. Find an upper limit of sample size that would ensure that this estimate does not deviate from the true value by more than 0.04 at 99% level of confidence.
32. In order to test the durability of a new paint, a highway department had test strips painted across heavily travelled roads in 15 different locations. If on the average, the test strips disappeared after they had been crossed by 1,46, 692 cars and with standard deviation 14,380 cars, construct 99% confidence interval for the true average number of cars it will take to wear off.

33. A machine is supposed to drill holes with a diameter of 1 inch. In fact, the diameters are normally distributed with a mean of 1.01 inches and a standard deviation of 0.02 inch. If there is a tolerance of 0.02 inch, the holes should be between 0.91 and 1.02 inches. What percentage of the holes drilled are within clearance ?
34. A machine is producing ball bearings with diameter of 0.5 inches. It is known that the standard deviation of the ball bearings is 0.05 inches. A sample of 100 ball bearings is selected and their average diameter is found to be 0.498 inches. Determine the 99 per cent confidence interval.
35. With a sample size of 400, the calculated standard error of mean is 2 with a mean of 120. What sample size would be required so that we could be 95% confident that the population mean is within ± 3.5 of the sample mean ?
36. A random sample of 160 people is taken and 120 were in favour of liberalising licensing regulations. With 95% confidence, what proportion of all people are in favour ?
37. A sample of 64 men from a population with known standard deviation of height of 2.4 inches gives a mean height of 69.8 inches. Find a 90% confidence interval of μ, the mean height of the men in the population. [69.3065 to 70.2935]
38. Given a population with a standard deviation of 8.6, What sample size is needed to estimate the mean of the population within ±0.5 of the sample mean with 99 per cent confidence ? *(MBA, KU, 2003)*

*****

15

Tests of Hypothesis

INTRODUCTION

A hypothesis is an assumption about the population parameter to be tested based on sample information. The statistical testing of hypothesis is the most important technique in statistical inference. Hypothesis tests are widely used in business and industry for making decisions. It is here that probability and sampling theory plays an ever increasing role in constructing the criteria on which business decisions are made. Very often in practice we are called upon to make decisions about population on the basis of sample information. For example, we may wish to decide on the basis of sample data whether a new medicine is really effective in curing a disease, whether one training procedure is better than another, etc. Such decisions are called *statistical decisions*.

In attempting to reach decisions, it is useful to make assumptions or guesses about the populations involved. Such assumptions, which may or may not be true, are called *statistical hypothesis* and in general are statements about the probability distributions of the population. The hypothesis is made about the value of some parameter, but the only facts available to estimate the true parameter are those provided by a sample. If the sample statistic differs from the hypothesis made about the population parameter, a decision must be made as to whether or not this difference is significant. If it is, the hypothesis is rejected. If not, it must be accepted. Hence, the term "tests of hypothesis".

Now, if θ be the parameter of the population and is the estimate of θ in the random sample drawn from the population, then the difference between θ and $\hat{\theta}$ should be small. In fact, there will be some difference between θ and $\hat{\theta}$ because $\hat{\theta}$ is based on sample observations and is different for different samples. Such a difference is known as difference due to sampling fluctuations. If the difference between θ and $\hat{\theta}$ is large, then the probability that it is exclusively due to sampling fluctuations is small. Difference which is caused because of sampling fluctuations is called insignificant difference and the difference due to some other reasons is known as significant difference. A significant difference arises due to the fact that either the sampling procedure is not purely random or sample is not from the given population.

Procedure of Hypothesis Testing

The general procedure followed in testing hypothesis comprises the following steps :

(1) *Set up a hypothesis*. The first step in hypothesis testing is to establish the hypothesis to be tested. Since statistical hypothesis are usually assumptions about the value of some unknown parameter, the hypothesis specifies a numerical value or range of values for the parameter. The conventional approach to hypothesis testing is not to construct single hypothesis about the population parameter, but rather to set up two different hypothesis. These hypothesis are normally referred to as (*i*) null hypothesis denoted by H_0, and (*ii*) alternative hypothesis denoted by H_1.

The null hypothesis asserts that there is no true difference in the sample statistic and population parameter under consideration (hence the word "null" which means invalid, void or amounting to nothing) and that the difference found is accidental arising out of fluctuations of sampling.

A hypothesis which states that there is no difference between assumed and actual value of the parameter is the null hypothesis and the hypothesis that is different from the null hypothesis is the alternative hypothesis. If the sample information leads us to reject H_0, then we will accept the alternative hypothesis H_1. Thus, the two hypothesis are constructed so that if one is true, the other is false and *vice versa*.

The rejection of the null hypothesis indicates that the differences have statistical significance and the acceptance of the null hypothesis indicates that the differences are due to chance. As against the null hypothesis, the alternative hypothesis specifies those values that the researcher believes to hold true. The alternative hypothesis may embrace the whole range of values rather than single point.

(2) *Set up a suitable significance level.* Having set up a hypothesis, the next step is to select a suitable level of significance. The confidence with which an experimenter rejects or retains null hypothesis depends on the significance level adopted. The level of significance, usually denoted by "α", is generally specified before any samples are drawn, so that results obtained will not influence our choice. Though any level of significance can be adopted, in practice, we either take 5 per cent or 1 per cent level of significance. When we take 5 per cent level of significance then there are about 5 chances out of 100 that we would reject the null hypothesis when it should be accepted, *i.e.*, we are about 95% confident that we have made the right decision. When we test a hypothesis at a 1 per cent level of significance, there is only one chance out of 100 that we would reject the null hypothesis when it should be accepted, *i.e.*, we are about 99% confident that we have made the right decision. When the null hypothesis is rejected at $\alpha = 0.5$, the test result is said to be "significant". When the null hypothesis is rejected at $\alpha = 0.01$, the test result is said to be "highly significant".

(3) *Determination of a suitable test statistic.* The third step is to determine a suitable test statistic and its distribution. Many of the test statistics that we shall encounter will be of the following form :

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Hypothesised population parameter}}{\text{Standard error of the sample statistic}}$$

(4) *Determine the critical region.* It is important to specify, before the sample is taken, which values of the test statistic will lead to a rejection of H_0 and which lead to acceptance of H_0. The former is called the *critical region.* The value of α, the level of significance, indicates the importance that one attaches to the consequences associated with incorrectly rejecting H_0. It can be shown that when the level of significance is α, the optimal critical region for a two-sided test consists of that $\alpha/2$ per cent of the area in the right-hand tail of the distribution plus that $\alpha/2$ per cent in the left hand tail. Thus, establishing a critical region is similar to determining a $100\,(1-\alpha)\%$ confidence interval. In general, one uses a level of significance of $\alpha = 0.05$, indicating that one is willing to accept a 5 per cent chance of being wrong to reject H_0.

(5) *Doing computations.* The fifth step in testing hypothesis is the performance of various computations from a random sample of size n, necessary for the test statistic obtained in step (3). Then, we need to see whether sample result falls in the critical region or in the acceptance regions.

(6) *Making decisions.* Finally, we may draw statistical conclusions and the management may take decisions. A statistical decision or conclusion comprises either accepting the null hypothesis or rejecting it. The decision will depend on whether the computed value of the test criterion falls in the region of rejection or the region of acceptance. If the hypothesis is being tested at 5 per cent level of significance and the observed set of results has a probability less than 5 per cent, we reject the null hypothesis and the difference between the sample statistic and the hypothetical population parameter is considered to be significant. On the other hand, if the testing statistic falls in the region of non-rejection, the null hypothesis is accepted and the difference between the sample statistic and the hypothetical population parameter is not regarded as significant, *i.e.*, it can be explained by chance variations.

Type I and Type II Errors

When a statistical hypothesis is tested, there are four possible results :

(1) The hypothesis is true but our test rejects it.

(2) The hypothesis is false but our test accepts it.

(3) The hypothesis is true and our test accepts it.

(4) The hypothesis is false and our test rejects it.

Obviously, the first two possibilities lead to errors. If we reject a hypothesis when it should be accepted (possibility No. 1), we say that a *Type I error* has been made. On the other hand, if we accept a hypothesis when it should be rejected (possibility No. 2), we say that a *Type II error* has been made. In either case a wrong decision or error in judgment has occurred.

TWO KINDS OF ERRORS IN HYPOTHESIS TESTING

Decision	*Condition* H_0 : True	H_0 : False
Accept H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

The probability of committing a *type I error* is designated as "α" and is called the *level of significance*. Therefore,

$$\alpha = P_r \text{ [Type I error]}$$
$$= P_r \text{ [Rejecting } H_0/H_0 \text{ is true]}$$

must be the complement of

$$\therefore \quad (1-\alpha) = P_r \text{ [Accepting } H_0/H_0 \text{ is true].}$$

This probability $(1 - \alpha)$ corresponds to the concept of 100 $(1-\alpha)$% confidence interval. Our efforts would obviously be to have a small probability of making a type I error. Hence the objective is to construct the test to *minimise* α.

Similarly, the probability of committing a type II error is designated by β. Thus

$$\beta = P_r \text{ [Type II error]}$$
$$= P_r \text{ [Accepting } H_0/H_0 \text{ is false]}$$

and $$(1-\beta) = P_r \text{ [Rejecting } H_0/H_0 \text{ is false].}$$

This probability $(1 - \beta)$ is known as the *power* of a statistical test.

The following table gives the probabilities associated with each of the four cells shown in the previous table :

The decision is :	The null hypothesis is True	The null hypothesis is False
Accept H_0	$(1-\alpha)$ Confidence level	β
Reject H_0	α	$(1-\beta)$ Power of the test
Sum	1.00	1.00

Note that the probability of each decision outcome is a conditional probability and the elements in the same column sum to 1.0, since the events with which they are associated are complement. However, α and β are not independent of each other, nor are they independent of the sample size n. When n is fixed, if α is lowered then β normally rises and *vice versa*. If n is increased, it is possible for both α and β to decrease. Since, increasing the sample size involves money and time, therefore, one should decide how much additional money and time, he is willing to spare on increasing the sample size in order to reduce the size of α and β.

In order for any tests of hypothesis or rules of decisions to be good, they must be designed so as to minimise errors of decision. However, this is not a simple matter, since for a given sample size, an attempt to decrease one type of error is accompanied in general by an increase in other type of error. The probability of making type I error is fixed in advance by the choice of level of significance employed in the test. We can make the type I error as small as we please, by lowering the level of significance. But, by doing so, we increase the chance of accepting a false hypothesis, *i.e.*, of making a type II error. It follows that it is impossible to minimise both errors simultaneously. In the long run, errors of type I are perhaps more likely to prove serious in research programmes in social sciences than are errors of type II. In practice, one type of error may be more serious than the other and so a compromise should be reached in favour of limitations of the more serious error. The only way to reduce both types of error is to increase the sample size which may or may not be possible.

One-Tailed and Two-Tailed Tests

Basically, there are three kinds of problems of tests of hypothesis. They include :

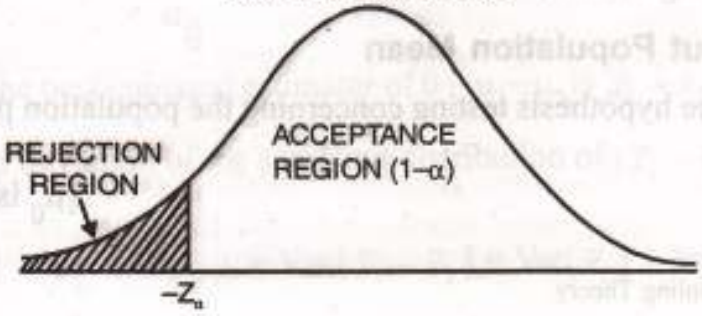
(*i*) two-tailed tests, (*ii*) right-tailed test, and (*iii*) left-tailed test.

Two-tailed test is that where the hypothesis about the population mean is rejected for value of falling into either tail of the sampling distribution. When the hypothesis about population mean is rejected only for value of falling into one of the tails of the sampling distribution, then it is known as one-tailed test. If, it is right tail then it is called right-tailed test or one-sided alternative to the right and if it is on the left tail, then, it is one-sided alternative to the left and called left-tailed test. For example, $H_0 : \mu = 100$ tested against $H_1 : \mu > 100$ or < 100 is one-tailed test since H_1 specifies that μ lies on particular side of 100. The same null hypothesis tested against $H_1 : \mu \neq 100$ is a two-tailed test since μ can be on either side of 100. The following diagrams would make it more clear :

TWO-TAILED TEST

RIGHT-TAILED TEST

LEFT-TAILED TEST



The following table gives critical values of z for both one-tailed and two-tailed tests at various levels of significance. Critical values of z for other levels of significance are found by use of the table of normal curve areas :

Level of Significance	0.10	0.05	0.01	0.005	0.0002
Critical value of z for one-tailed tests	– 1.28 or 1.28	– 1.645 or 1.645	– 2.33 or 2.33	– 2.58 or 2.58	– 2.88 or 2.88
Critical value of z for two-tailed tests	– 1.645 and 1.645	– 1.96 and 1.96	– 2.58 and 2.58	– 2.81 and 2.81	– 3.08 and 3.08

Tests of Hypothesis Concerning Large Samples

Though, it is difficult to draw a clear-cut line of demarcation between large and small samples, it is generally agreed that if the size of sample exceeds 30, it should be regarded as a large sample. The tests of significance used for large samples are different from the ones used for small samples* for the reason that the assumptions we make in case of large samples do not hold for small samples. Tests of hypothesis involving large samples are based on the following assumptions :

(1) The sampling distribution of a sample statistic is approximately normal.

(2) Values given by the samples are sufficiently close to the population value and can be used in its place for the standard error of the estimate.

Thus, we have seen that the normal distribution plays a vital role in tests of hypothesis based on large samples (central limit theorem).

Suppose $\hat{\theta}$ is an unbiased estimate of θ, the population parameter. On the basis of $\hat{\theta}$, taken from sample observations, it is to test the hypothesis whether the sample is drawn from a population whose parameter value is θ, *i.e.*, we have to test the hypothesis

$$H_0 : \theta = \hat{\theta}$$

If sampling distribution of θ is normal, then

$$z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0, 1).$$

Let us test the hypothesis at 100 α% level of significance. From tables of area under the standard normal curve corresponding to given α, we can find an ordinate z_α such that

$$P_r [\, | z_\alpha | > z_\alpha \,] = \alpha$$
$$P_r [\, -z_\alpha \leq z \leq z_\alpha \,] = 1 - \alpha.$$

If $\alpha = .01$, then $z_\alpha = 2.58$ and if $\alpha = 0.05$, then $z_\alpha = 1.96$, and so on.

If the difference between $\hat{\theta}$ and θ is more than z_α times, the standard error of $\hat{\theta}$, the difference is regarded significant and H_0 is rejected at 100 α% level of significance and if the difference between θ and $\hat{\theta}$ is less than or equal to z_α times the standard error of $\hat{\theta}$, the difference is insignificant and H_0 is accepted at 100α% level of significance.

Testing Hypothesis about Population Mean

(*a*) We shall first take the hypothesis testing concerning the population parameter μ by considering the two-tailed test.

$H_0 : \mu = \mu_0$ [μ_0 is hypothesised value of μ.]

*See Chapter 16 on Small Sampling Theory.

Since the best unbiased estimator of μ is the sample mean $\bar{x}$, therefore, we shall focus our attention on the sampling distribution of $\bar{x}$. From central limit theorem, we know

$$\bar{x} \sim N\left(\mu, \sigma_{\bar{x}}\right)$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

where

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \text{[If } \sigma \text{ is known.]}$$

$$= \frac{s}{\sqrt{n}} \qquad \text{[If } \sigma \text{ is unknown for large samples.]}$$

If the calculated value of $z < -z_{\alpha/2}$ or $> z_{\alpha/2}$, the null hypothesis is rejected.

(*b*) If the hypothesis involves a right-tailed test. For example,

$$H_0 : \mu \leq \mu_0 \text{ and } H_1 : \mu > \mu_0.$$

For the calculated value $z > z_{\alpha}$, the null hypothesis is rejected.

(*c*) If the hypothesis involves a left-tailed test, *i.e.*,

$$H_0 : \mu \geq \mu_0 \text{ and } H_1 : \mu < \mu_0$$

then for the value $z < -z_{\alpha}$, the null hypothesis is rejected.

Illustration 1. The mean lifetime of a sample of 100 light tubes produced by a company is found to be 1,580 hours with standard deviation of 90 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1,600 hours.

Solution. The null hypothesis is that there is no significant difference between the sample mean and hypothetical population mean, *i.e.*, $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

where

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \text{[Since } \sigma \text{ is unknown for large samples.]}$$

$$z = \frac{1580 - 1600}{90/\sqrt{100}} = -2.22$$

The critical value is $z = \pm 1.96$ for a two-tailed test at 5% level of significance. Since, the computed value of $z = -2.22$ falls in the rejection region, we reject the null hypothesis. Hence, the mean lifetime of the tubes produced by the company may not be 1,600 hours.

Testing Hypothesis about the Difference between Two Means

The test statistic for testing the difference between two population means, when the populations are normally distributed, is based on the general form of the standard normal statistic as given below :

$$z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

where $\theta = \mu_1 - \mu_2$. Since, the best unbiased estimator of $\theta = \mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$, therefore, $\hat{\theta}$ is replaced by $\bar{x}_1 - \bar{x}_2$. $\sigma_{\hat{\theta}}$, the standard deviation of the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is given by

$$\sigma_{\hat{\theta}}^2 = \sigma_{\bar{x}_1 - \bar{x}_2}^2 = \text{Var}\,[\bar{x}_1 - \bar{x}_2] = \text{Var}[\bar{x}_1] + \text{Var}\,[\bar{x}_2] = \frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}$$

Therefore, the z statistic is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$

Then, the z statistic is reduced to $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

At 5% level of significance, the critical value of z for two-tailed test = ± 1.96. If the computed value of z is greater than +1.96 or less than –1.96, then reject H_0, otherwise accept H_0.

In case σ_1^2 and σ_2^2 are not known then for large samples, s_1^2 and s_2^2 can be used instead.

Illustration 2. You are working as a purchase manager for a company. The following information has been supplied to you by two manufacturers of electric bulbs :

	Company *A*	Company *B*
Mean life (in hours)	1,300	1,288
Standard deviation (in hours)	82	93
Sample size	100	100

Which brand of bulbs are you going to purchase if you desire to take a risk of 5% ? *(MBA, Kumaun Univ., 2002)*

Solution. Let us take the null hypothesis that there is no significant difference in the quality of the two brands of bulbs, *i.e.*, $H_0 : \mu_1 = \mu_2$

[Since σ_1^2 and σ_2^2 are not known, therefore, can be replaced by s_1^2 and s_2^2.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1300 - 1288}{\sqrt{\frac{(82)^2}{100} + \frac{(93)^2}{100}}}$$

$$= \frac{12}{\sqrt{67.24 + 86.49}} = \frac{12}{12.399} = 0.968$$

Since our computed value of $z = 0.968$ is less than critical value of $z = 1.96$ (5% level), we accept the null hypothesis. Hence, the quality of two brands of bulbs do not differ significantly.

Test of Hypothesis Concerning Attributes

As distinguished from variables where quantitative measurement of a phenomenon is possible in case of attributes we can only find out the presence or absence of a certain characteristic. For example, in the study of attribute 'employment' a sample may be taken and people classified as employed and unemployed. With such data, the binomial type of problem may be formed. The selection of an individual on sampling may be called 'event', the appearance of an attribute '*A*' may be taken as "success" and its non-appearance, as "failure". The sampling distribution of the number of successes, being a binomial probability model would have its mean $\mu = np$ and its standard deviation $\sigma = \sqrt{npq}$.

Then $$z = \frac{x - np}{\sqrt{npq}} \sim N(0, 1).$$

Illustration 3. In 600 throws of six-faced die, odd points appeared 360 times. Would you say that the die is fair at 5% level of significance ?

Solution. Let us take the hypothesis that the die is not biased.

$$p = q = \frac{1}{2},\ n = 600,\ np = 300.$$

Applying the formula ;

$$z = \frac{x - np}{\sqrt{npq}} = \frac{360 - 300}{\sqrt{600 \times \frac{1}{2} \times \frac{1}{2}}} = \frac{60}{12.25} = 4.9.$$

Since, the computed value of z is greater than the table value (1.96 at 5% level of significance), the hypothesis is rejected. Hence, the die does not seem to be fair.

Testing Hypothesis about a Population Proportion

The population parameter of interest is population proportion π. If the sample size is large, then sample proportion p will be approximately normally distributed. Then

$$z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0, 1).$$

The null hypothesis is that there is no significant difference between the sample proportion and population proportion, *i.e.*, $H_0 : p = \pi$

Since the sample proportion p is unbiased estimator of π,

$$\therefore \qquad z = \frac{p - \pi}{\sigma_p};\ \text{where } \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Therefore, the statistic

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1).$$

If $|z| < z_\alpha$, the null hypothesis is rejected with $100\alpha\%$ level of significance.

Illustration 4. A sales clerk in the departmental store claims that 60% of the shoppers entering the store leave without making a purchase. A random sample of 50 shoppers showed that 35 of them left without buying anything. Are these sample results consistent with the claim of the sales clerk ? Use a level of significance of 0.05.

Solution. The null hypothesis is

$$H_0 : \pi = 0.60.$$

The sample proportion $p = \frac{35}{50} = 0.70.$

Using the z statistic, we have

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.70 - 0.60}{\sqrt{(0.6)(0.4)/50}} = 1.45.$$

The critical value of z is 1.64 at 5% level of significance.

Since, the computed value of $z = 1.45$ is less than the critical value of $z = 1.64$, therefore, the null hypothesis cannot be rejected. Hence, based on this sample data, we cannot reject the claim of the sales clerk.

Testing Hypothesis about the Difference Between Two Proportions

Let p_1 and p_2 be the sample proportions obtained in large samples of sizes n_1 and n_2 drawn from respective populations having proportions π_1 and π_2. We can test the null hypothesis that there is no

difference between the population proportions, *i.e.*,

$$H_0 : \pi_1 = \pi_2.$$

As shown in the earlier chapter, the sampling distribution of differences in proportion, $p_1 - p_2$ is normally distributed with mean

$$\mu_{p_1 - p_2} = \pi_1 - \pi_2$$

and standard deviation

$$\sigma_{p_1 - p_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}.$$

Therefore, the statistic is

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$$

If the null hypothesis is true, p_1 and p_2 are two independent unbiased estimators of the same parameter $\pi_1 = \pi_2 = \pi$. Thus, our procedure is to pool our observations to obtain the best estimate of the common value π. The pooled estimate of π is the weighted mean of the two sample proportions, *i.e.*,

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Our test statistic then becomes

$$z = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}}, \quad \text{where } \sigma_{p_1 - p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Illustration 5. In a random sample of 100 persons taken from village *A*, 60 are found to be consuming tea. In another sample of 200 persons taken from village *B*, 100 persons are found to be consuming tea. Do the data reveal significant difference between the two villages so far as the habit of taking tea is concerned ? *(MBA, Delhi Univ., 1999)*

Solution. Let us take the hypothesis that there is no significant difference between the two villages so far as the habit of taking tea is concerned, *i.e.*, $\pi_1 = \pi_2$.

We are given :

$$p_1 = \frac{x_1}{n_1} = \frac{60}{100} = 0.6, n_1 = 100.$$

$$p_2 = \frac{x_2}{n_2} = \frac{100}{200} = 0.5, n_2 = 200.$$

The appropriate statistics to be used here is given by

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{60 + 100}{100 + 200} = 0.53$$

$$z = \frac{0.6 - 0.5}{\sqrt{(0.53)(0.47)\left(\frac{1}{100} + \frac{1}{200}\right)}} = \frac{0.1}{\sqrt{(0.53)(0.47)(0.015)}}$$

$$= \frac{0.1}{\sqrt{0.0037}} = \frac{0.1}{0.0608} = 1.64.$$

Since, the computed value of z is less than the critical value of $z = 1.96$ at 5% level of significance, therefore, we accept the hypothesis. Hence, we conclude that there is no significant difference in the habit of taking tea in the two villages A and B.

Illustration 6. Before an increase in excise duty on tea, 400 people out of a sample of 500 people were found to be tea drinkers. After an increase in duty, 400 people were tea drinkers in a sample of 600 people. State, whether there is a significant decrease in the consumption of tea. *(MBA, Delhi Univ., 2002)*

Solution. Let us take the hypothesis that there is no significant decrease in the consumption of tea after the increase in duty, *i.e.*, $\pi_1 = \pi_2$.

We are given
$$p_1 = \frac{x_1}{n_1} = \frac{400}{500} = 0.8, n_1 = 500.$$
$$p_2 = \frac{x_2}{n_2} = \frac{400}{600} = 0.667, n_2 = 600.$$

The appropriate test statistic to be used here is given by

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where
$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{400 + 400}{500 + 600} = 0.73$$
$$z = \frac{0.8 - 0.667}{\sqrt{(0.73)(0.27)\left(\frac{1}{500} + \frac{1}{600}\right)}} = \frac{0.133}{\sqrt{(0.73)(0.27)(0.0037)}}$$
$$= \frac{0.133}{\sqrt{0.00073}} = \frac{0.133}{0.027} = 4.93.$$

Since, the computed value of z is greater than the critical value of $z = 1.96$ at 5% level of significance, therefore, hypothesis is rejected. Hence, there is a significant decrease in the consumption of tea after an increase in duty.

MISCELLANEOUS ILLUSTRATIONS

Illustration 7. From the following data obtained from a sample of 1,000 persons, calculate the standard error of mean :

Weekly Earnings (Rs. hundred) :	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
No. of persons :	50	100	150	200	200	100	100	100

Is it likely that the sample has come fom the population with an average weekly earnings of Rs. 4,200.

Solution.

CALCULATION OF STANDARD DEVIATION

Weekly Earnings (Rs. hundred)	X	f	$(X-45)/10$ d	fd	fd^2
0–10	5	50	–4	–200	800
10–20	15	100	–3	–300	900
20–30	25	150	–2	–300	600
30–40	35	200	–1	–200	200
40–50	45	200	0	0	0
50–60	55	100	+1	+100	100
60–70	65	100	+2	+200	400
70–80	75	100	+3	+300	900
		$n = 1000$		$\Sigma fd = -400$	$\Sigma fd^2 = 3,900$

$$\bar{x} = A + \frac{\Sigma fd}{n} \times i = 45 - \frac{400}{1000} \times 10 = 41$$

$$s = \sqrt{\frac{\Sigma fd^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2} \times i = \sqrt{\frac{3900}{1000} - \left(\frac{-400}{1000}\right)^2} \times 10$$
$$= 1.934 \times 10 = 19.34$$
$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{19.34}{\sqrt{1000}} = \frac{19.34}{31.62} = 0.612$$

Therefore, the standard error of mean is 0.612.

$$\bar{x} = 41, \mu = 42, \sigma_{\bar{x}} = 0.612.$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{41-42}{0.612} = -1.634$$

Since, the computed value of z is less than the critical value of $z = \pm 1.96$, it is not significant and hence there is no significant difference between the sample average and the population average weekly earnings and the difference could have arisen due to fluctuations of sampling.

Illustration 8. A sample of 400 managers is found to have a mean height of 171.38 cms. Can it be reasonably regarded as a sample from a large population of mean height 171.17 cms and standard deviation of 3.30 cms ?

Solution. The null hypothesis is that there is no significant difference between the sample mean height and the population mean height.

Given $\bar{x} = 171.38, \mu = 171.17, n = 400$, and $\sigma = 3.30$

Applying the test statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{171.38-171.17}{3.30/\sqrt{400}} = \frac{0.21}{0.16} = 1.31$$

Since, the computed value of $z = 1.31$ is less than critical value of $z = 1.96$ at 5% level of significance, therefore, the null hypothesis is accepted. Hence, there is no significant difference between the sample mean height and population mean height.

Illustration 9. Intelligence test given to two groups of boys and girls gave the following information :

	Mean Score	*S.D.*	*Number*
Girls	75	10	50
Boys	70	12	100

Is the difference in the mean scores of boys and girls statistically significant ? ***(MBA, S.V. Univ., 2004; MBA, DU,2005)***

Solution. Let us take the hypothesis that the difference in the mean score of boys and girls is not significant, *i.e.*, $\mu_1 = \mu_2$

We are given $\bar{x}_1 = 75, \bar{x}_2 = 70, s_1^2 = 100, s_2^2 = 144, n_1 = 50, n_2 = 100.$

The appropriate statistic to be used here is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\left[\text{Since } \sigma_1^2 = s_1^2; \sigma_2^2 = s_2^2 \text{ and } \mu_1 = \mu_2\right]$$

$$= \frac{75-70}{\sqrt{\frac{100}{50} + \frac{144}{100}}} = \frac{5}{\sqrt{3.44}} = \frac{5}{1.855} = 2.695$$

Since, the computed value $z = 2.695$ is greater than the critical value of $z = 2.58$ at 1% level of significance, therefore, the hypothesis is rejected. Hence, the difference in the mean score of boys and girls is statistically significant.

Illustration 10. In a survey of buying habits, 400 women shoppers are chosen at random in super market *A*. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For another group of 400 women shoppers chosen at random in super market *B* located in another area of the same 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 expenditures of the populations of women shoppers are equal.

Solution. The null hypothesis is that the average weekly food expenditures of the two populations are same, *i.e.*, $\mu_1 = \mu_2$.

Since σ_1^2 and σ_2^2 (the population variances) are not known, we can estimate from the sample variances (provided sample size is large), *i.e.*,

$$\sigma_1^2 = s_1^2, \quad \sigma_2^2 = s_2^2$$

Given : $n_1 = 400, n_2 = 400, \bar{x}_1 = 250, \bar{x}_2 = 220, s_1 = 40, s_2 = 55$

Applying the test statistic

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad [\text{since } \sigma_1^2 = s_1^2; \sigma_2^2 = s_2^2]$$

$$= \frac{250-220}{\sqrt{\frac{(40)^2}{400}+\frac{(55)^2}{400}}} = \frac{30 \times 20}{\sqrt{4625}} = \frac{600}{68.01} = 8.822$$

Since, the value of z is much greater than 3, the null hypothesis is rejected. Hence, the average weekly expenditure of two populations of women shoppers differ significantly.

Illustration 11. A dice is thrown 49152 times and of these 25145 yielded either 4 or 5 or 6. Is this consistent with the hypothesis that the dice must be unbiased ?

Solution. Let the coming of 4, 5 or 6 be termed as success, then the null hypothesis can be stated as that the dice is unbiased.

Given, $n = 49152$, and p = proportion of success $= \frac{25145}{49152} = 0.512$

The appropriate statistic to be used is

$$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{p-\pi}{\sqrt{\frac{p(1-p)}{n}}}$$

$$= \frac{0.512-0.5}{\sqrt{\frac{(0.512)(0.488)}{49152}}} = \frac{0.012}{0.002} = 6.0$$

Since, the computed value of $z = 6$ is much greater than the critical value of $z = 3$, it is significant, and therefore, null hypothesis is rejected. Hence, the dice is certainly biased.

Illustration 12. 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 then timed on 50 emergency calls, getting a mean of 9.3 minutes with a standard deviation of 1.8 minutes. At the level of significance of 0.05, does this constitute evidence that the figure claimed is too low ?

Solution. Let us take the hypothesis that there is no significant difference between the figure observed and the figure claimed, *i.e.*, 9.3 and 8.9.

We are given : $\mu = 8.9$, $\bar{x} = 9.3$, $s = 1.8$, $n = 50$.

The appropriate statistic to be used is given by

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \qquad [\sigma^2 = s^2 \text{ for large samples}]$$

$$= \frac{9.3-8.9}{1.8/\sqrt{50}} = \frac{0.4}{0.25} = 1.6$$

Since, the computed value of $z = 1.6$ is less than the critical value of $z = 1.96$ at 5% level of significance, therefore, the hypothesis is accepted. Hence, there is no significant difference between the average figure observed and the average figure claimed.

Illustration 13. A coin is tossed 100 times under identical conditions independently yielding 30 heads and 70 tails. Test at 1% level of significance, whether or not the coin is unbiased. State clearly the null hypothesis and the alternative hypothesis.

Solution. Let the null hypothesis be that the coin is unbiased. If p is the probability of getting head, then

$$H_0 : \pi = 0.5 \text{ and } H_1 : \pi \neq 0.5.$$

The appropriate statistic to be used here is z-statistic. We are given :

$p = 0.3$, $\pi = 0.5$ and $n = 100$

$$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.3-0.5}{\sqrt{\frac{(0.5)(0.5)}{100}}} = \frac{-0.2 \times 10}{\sqrt{0.25}} = -4.$$

Since, the computed value of $z = -4$ is greater than critical values of $z = \pm 2.58$ at 1% level of significance, therefore, we reject the null hypothesis. Hence, the coin is biased.

Illustration 14. A product is produced in two ways. A pilot test on 64 times from each method indicates that the product of Method 1 has sample mean tensile strength 106 lbs and a standard deviation 12 lbs, whereas in Method 2 the corresponding values of mean and standard deviation are 100 lbs and 10 lbs respectively. Greater tensile strength in the product is preferable. Use an appropriate large sample test at 5% level of significance to test whether or not Method 1 is better for processing the product. State clearly the null hypothesis.

(MBA, Delhi Univ., 2003)

Solution. Let the null hypothesis be that there is no significant difference between Method 1 and Method 2, *i.e.*, $H_0 : \mu_1 = \mu_2$

We are given

$$\bar{x}_1 = 106,\ \bar{x}_2 = 106, s_1 = 12, s_2 = 10, n_1 = n_2 = 64$$

Using the test statistic $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{106-100}{\sqrt{\dfrac{(12)^2}{64} + \dfrac{(10)^2}{64}}} = \frac{6 \times 8}{\sqrt{244}} = 3.07.$$

Since, the computed value of $z = 3.07$ is greater than the critical value of $z = 1.64$ at 5% level of significance, the null hypothesis is rejected. Hence, Method 1 is better than Method 2.

Illustration 15. A company is considering two different television advertisements for the promotion of a new product. Management believes that advertisement *A* is more effective than advertisement *B*. Two test market areas with virtually identical consumer characteristics are selected : advertisement *A* is used in one area and advertisement *B* in the other area. In a random sample of 60 customers who saw advertisement *A*, 18 tried the product. In a random sample of 100 customers who saw advertisement *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 ? *(MBA, IGNOU 2002; MBA, Delhi Univ., 2005)*

Solution. Let the null hypothesis be that there is no significant difference in the effectiveness of the two advertisements *A* and *B*, *i.e.*, $H_0 : \pi_1 = \pi_2$.

The appropriate statistic to be used is

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sigma_{p_1 - p_2}} = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \quad [\because \pi_1 = \pi_2]$$

where $p_1 = \dfrac{x_1}{n_1} = \dfrac{18}{60} = 0.30,\ n_1 = 60$

$$p_2 = \frac{x_2}{n_2} = \frac{22}{100} = 0.22,\ n_2 = 100$$

and $p = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{40}{160} = 0.25$

$$z = \frac{0.30 - 0.22}{\sqrt{(0.25)(0.75)\left(\dfrac{1}{60} + \dfrac{1}{100}\right)}} = \frac{0.08}{\sqrt{0.005}} = \frac{0.08}{0.071} = 1.13$$

Since, the computed value of $z = 1.13$ is less than the critical value of $z = 1.645$* at 5% level of significance, therefore, the null hypothesis is accepted. Hence, there is no significant difference in the effectiveness of the two advertisements *A* and *B*.

Illustration 16. 500 units from a factory are inspected and 12 are found to be defective, 800 units from another factory are inspected and 12 are found to be defective. Can it be concluded at 5% level of significance that production at second factory is better than in first factory ? *(MBA, Delhi Univ., 2007)*

Solution. Let us, take the null hypothesis that there is no significant difference in the proportion of defective items in the two factories.

$$p_1 = \frac{x_1}{n_1} = \frac{12}{500} = 0.024;\ p_2 = \frac{x_2}{n_2} = \frac{12}{800} = 0.015$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

* Normally in testing hypothesis, we use two-tailed test and the critical value of z at 5% level is 1.96. In this question, one-tailed test has been used and the critical value at 5% is 1.645.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{12+12}{500+800} = 0.018$$

$$z = \frac{0.024 - 0.015}{\sqrt{(0.018)\,(0.982)\,(0.00325)}} = \frac{0.009}{0.0076} = 1.184.$$

Since, the computed value of z is less than the critical value of $z = 1.96$ at 5% level of significance, therefore, our null hypothesis holds good. Hence, we cannot conclude that the production in the second factory is better than in the first factory.

Illustration 17. A buyer of electric bulbs bought 100 bulbs each of two famous brands. Upon testing these he found that brand A had a mean life of 1500 hours with a standard deviation of 50 hours whereas brand B had a mean life of 1530 hours with a standard deviation of 60 hours. Can it be concluded at 5 per cent level of significance that the two brands differ significantly in quality of the bulbs.

Solution. Let us, take the null hypothesis that the two brands of bulbs do not differ significantly in quality.

We are given $\bar{x}_1 = 1500$, $\bar{x}_2 = 1530$, $s_1 = 50$, $s_2 = 60$, $n_1 = 100$, $n_2 = 100$.

The appropriate statistic to be used here is given by :

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1500 - 1530}{\sqrt{\frac{(50)^2}{100} + \frac{(60)^2}{100}}} = -\frac{30}{7.81} = -3.84.$$

Since, the computed value of z is more than the table value of $z = 1.96$ at 5% level of significance, the null hypothesis is rejected. Hence, the two brands of bulbs differ significantly in quality.

Illustration 18. Two types of new cars produced in India are tested for petrol mileage. One group consisting of 36 cars averaged 14 kms. per litre. While the other group consisting of 72 cars averaged 12.5 kms. per litre.

(*a*) What test statistic is appropriate, if

$$\sigma_1^2 = 1.5 \text{ and } \sigma_2^2 = 2.0 \text{ ?}$$

(*b*) Test, whether there exists a significant difference in the petrol consumption of these two types of cars (use $\alpha = 0.01$).

(MBA, IIT Roorkee, 2000)

Solution. We are given the following information :

$$n_1 = 36 \qquad \bar{x}_1 = 14 \qquad \sigma_1^2 = 1.5$$

$$n_2 = 72 \qquad \bar{x}_2 = 12.5 \qquad \sigma_2^2 = 2.0$$

(*a*) The appropriate test statistic to be used is the test of difference between two means.

(*b*) Let us take the null hypothesis that there is no significant difference in the petrol consumption of the two types of cars, *i.e.*, $H_0 : \mu_1 = \mu_2$.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{14 - 12.5}{\sqrt{\frac{1.5}{36} + \frac{2}{72}}} = \frac{1.5}{0.264} = 5.68$$

Since, the calculated value of $z = 5.68$ is greater than the critical value of $z = 2.58$ (1% level), the null hypothesis is rejected. Hence, there is a significant difference in the petrol consumption of the two types of cars.

Illustration 19. The Educational Testing Service conducted a study to investigate difference between the scores of male and female students on the Scholastic Aptitude Test. The study identified a random sample of 562 female and 852 male students who had achieved the same high score on the mathematics portion of the test. That is, the female and male students were viewed as having similarly high abilities in mathematics. The verbal scores for the two samples are as given :

Female students : $\bar{x}_1 = 547$; $s_1 = 83$; Male student : $\bar{x}_2 = 525$; $s_2 = 78$

Do the data support the conclusion that given a population of female students and a population of male students with similarly high mathematics abilities, the female students will have a significantly higher verbal ability ? Test at a 5% level of significance. What is your conclusion ?

(MBA, DU, 2003)

Solution : Given : $\bar{x}_1 = 547$, $\bar{x}_2 = 525$, $s_1 = 83$, $s_2 = 78$, $n_1 = 562$, $n_2 = 852$

Let us take the null hypothesis be that there is no significant difference between male and female verbal ability, *i.e.*, $H_0 : \mu_1 - \mu_2 \geq 0$ and $H_1 : \mu_1 - \mu_2 < 0$

Using $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{547 - 525}{\sqrt{\frac{(83)^2}{562} + \frac{(78)^2}{852}}}$$

$$= \frac{22}{\sqrt{\frac{6889}{562} + \frac{6084}{852}}} = \frac{22}{\sqrt{12.258 + 7.1408}} = \frac{22}{\sqrt{19.3988}} = \frac{22}{4.044} = 4.995$$

The computed value of z is greater than the table value of $z = \pm 1.96$. Therefore, reject the null hypothesis. Hence, there is a significant difference between the male and female verbal ability or female student have higher verbal ability.

Illustration 20. Record of several years of applicants for admission at FMS showed their mean score is 315. An administrator is interested in knowing whether the caliber of recent applicants has changed. For testing this hypothesis the scores of a sample of 100 applicants from the scores of recent applicants is obtained from admission office. The mean for this turned out to be 328. The sample standard deviation is 38, which may also be assumed for the population. Test the hypothesis using 5% level of significance. *(MBA, Delhi Univ., 2009)*

Solution. Let us take the hypothesis that there is no change in the calibre of recent applicants. The appropriate statistic to be used is given by

$$z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$$

$\bar{X} = 315, \mu = 328, \sigma = 38, n = 100$

Substituting the values

$$z = \frac{315 - 328}{38} \times \sqrt{100} = \frac{13}{38} \times 10 = \frac{130}{38} = 3.421$$

Since the calculated value of z is more than the critical value of $z = 1.96$ at 5% level of significance, the hypothesis is rejected. Hence, there seems to be a change in the cablire of recent applicants.

PROBLEMS

1-A: Answer the following questions, each question carries **one** mark:

(*i*) What is a hypothesis ?
(*ii*) What is Null Hypothesis ? *(MBA, Madurai-Kamaraj Univ., 2003)*
(*iii*) What is standard error ?
(*iv*) Explain clearly the terms "standard error" and "sampling disribution". *(MBA, Madurai-Kamaraj Univ., 2008)*
(*v*) What is type I error ?
(*vi*) What is type II error ?
(*vii*) Differentiate between type I and type II error ?
(*viii*) What are the critical values at 1% and 5% level.
(*ix*) What is degrees of freedom ?
(*x*) What do you understand by large sample ?

1-B : Answer the following questions, each question carries marks **four** :

(*i*) State the procedure followed in testing of hypothesis. *(M.Com., M.K. Univ., 2003)*
(*ii*) Differentiate the following pairs of concepts :
  (*i*) Statistic and parameter.
  (*ii*) Null and Alternative Hypothesis.
  (*iii*) Type I and Type II error.
(*iii*) Explain the different steps in testing hypothesis.
(*iv*) Define the two types of errors in testing a statistical hypothesis. *(MBA, Anna Univ., 2003)*
(*v*) Explain the difference between two proportions in test of hypothesis.

2. What is test of hypothesis ? Discuss various tests of hypothesis for the cases when the size of sample is large.
3. Explain the procedure generally followed in testing of a hypothesis.
4. Describe the various steps involved in testing of hypothesis. What is the role of standard error in testing of hypothesis ? *(M.Com., Delhi Univ.; MBA UP Tech. Univ., 2007)*
5. Define the standard error of a statistic. How is it helpful in testing of hypothesis and decision-making ?
6. What do you understand by null hypothesis and level of significance ? Explain with the help of an example.
7. Write short notes on the following :
   (*i*) Type I and Type II error.
   (*ii*) In the hypothesis testing process, what is the importance of *null hypothesis ?*
   (*iii*) "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 ?

8. Explain clearly the procedure of testing hypothesis. Also point out the assumptions in hypothesis testing in large samples.
(M.Phil., Kurukshetra Univ.)

9. Differentiate the following pairs of concepts :
   (*i*) Statistic and parameter.
   (*ii*) Critical region and acceptance region.
   (*iii*) Null and alternative hypothesis.
   (*iv*) One-tailed and two-tailed test.
   (*v*) Type I and Type II errors.

10. There is always a trade off between Type I and Type II errors. Discuss.

11. Intelligence test on two groups of boys and girls gave the following results :

	Mean	*S.D.*	*Sample size*
Girls :	75	15	150
Boys :	70	20	250

Is there a significant difference in the mean scores obtained by boys and girls ?
(M.Com., Madurai-Kamaraj Univ., 2002; MBA, Kumaun Univ., 2009)

12. In a sample of 1000 persons from the village of Himachal Pradesh, 660 are found to be consumers of rice and the rest consumers of wheat. Can it be concluded that both the food articles are equally popular ?

13. Random samples of 100 bolts manufactured by machine '*A*' and 50 bolts from machine '*B*' showed 10 and 6 defective bolts respectively. Is there a significant difference in the performance of the two machines ?

14. The mean lifetime of 200 fluorescent light tubes made by a company gave mean lifetime of 1560 hours with a standard deviation of 50 hours. Is it likely that the sample has come from a population with a mean lifetime of 1,500 hours ?

15. A soap manufacturer wanted to know what percentage of the citizens of Mumbai use his soap. He conducted a survey and found that out of 500 persons selected at random for the purpose, only 10% use his soap. He spent Rs. 5 lakh on an advertisement campaign to attract more customers. In order to know the result of his campaign he conducted another survey and found that out of 600 persons 15% are using his soap. Do you think that the expenditure has really increased the percentage of citizens of Mumbai using his soap ?

16. A machine puts out 20 imperfect articles in a sample of 1000. After the machine is overhauled, it puts out 5 imperfect articles in a sample of 300. Has the machine improved ?

17. In North Delhi, out of a random sample of 500 households, 25% declared that they were regular readers of 'Femina'. In South Delhi, the proportion in a sample of 600 was 30%. Is there a significant difference in the two proportions ?

18. A firm found with the help of a sample survey of a city (size of a sample 900) the 3/4ths of the population consumes things produced by them. The firm then advertised the goods in paper and on radio. After one year, a sample of size 1000 reveals that proportions of consumers of the goods produced by the firm is 4/5th. Is this rise significant to indicate that the advertisement was effective ?

19. X is a normally distributed random variable. The variance of X is σ^2 and is known. Construct a test criterion to test the hypothesis that the mean of X is equal to μ_0 (a given constant). Suppose σ^2 was unknown, suggest an unbiased estimator of σ^2 and give (state) the test criterion to be used in this case.

20. A sample of size 400 was drawn and the sample mean was found to be 99. Test whether this sample could have come from a normal population with mean 100 and variance 64 at 5% level of significance.

21. A manufacturer claimed that at least 95% of the equipments which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at a significant level (*i*) .05; (*ii*) 0.1.

22. 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 gm with a standard deviation of 12 gm while the corresponding figure in a sample of 400 items from the other process are 124 and 14 gms. Is the difference between the mean weights significant at 1% level of significance ?

23. The mean breaking strength of the 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 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 1% level of significance ?

24. A sample of 400 male students is found to have a mean height of 171.38 cm. Can it be reasonably regarded and a sample from a large population with mean height 171.17 cm and standard deviation 3.30 cms ?

25. Give the requirements for applying Normal distribution to a problem of testing the significance of single mean. Give the null hypothesis H_0 and describe the procedure of testing H_0 against various possible alternative hypothesis H_1 at 5% level of significance.

Given $\bar{x}_1 = 82$, $\sigma = 10$, $n = 100$, test the hypothesis that $\mu = 86$.

26. In a random sample of 500 persons from town *A*, 200 are found to be consumers of wheat. In a sample of 400 from town *B*, 220 are found to be consumers of wheat. Do these data reveal a significant difference between town *A* and town *B* so far as the proportion of wheat consumer is concerned ?

27. A company produces two makes of bulbs, *A* and *B*. 200 bulbs of each make were tested and it was found that make *A* has mean life of 2560 hours and S.D. 90 hours, whereas make *B* had 2650 hours mean life and S.D. 75 hours. Is there a significant difference between the mean life of two makes ?

28. An equal opportunities committee is conducting an investigation if in comparable jobs, men and women workers are paid identical wages. The following information is obtained on 75 males and 64 females :

Salary	Male	Female
Mean (Rs.)	11,530	10,620
S.D.	780	750

Test at 5% level of significance, whether men and women workers are paid identical wages.

29. In a credit co-operative run by a large company it was found that during the past year, a sample of 300 loans issued showed that 37% of the loans were made to women employees. A similar study carried out 5 years ago showed that the proportion of women employees seeking loans was 32%. Do these data give sufficient evidence to conclude that more women employees are seeking loans in the recent year than before.

Use a 5% significance level for test.

30. Data were collected from two cities as regards the starting stipend paid to new management trainees.

Do the data give evidence that the stipend paid in city *B* is significantly more than that in city *A*?

Test at a significance level of 1%.

City	*Monthly Stipend (Means)*	*Sample Standard Deviations*	*Sample Size*
A	Rs. 8,400	Rs. 80	200
B	Rs. 8,600	Rs. 120	175

31. A manufacturer of steel rods considers that the manufacturing is working properly, if the mean length of the rods is 8.6 inches. The standard deviation of these rods always runs about 0.3 inch. The manufacturer would like to see, if the process is working correctly by taking a random sample of size $n = 36$. There is no indication whether or not the rods may be too short or too long.

(*a*) Establish null and alternative hypothesis for this problem.

(*b*) Would you use a one-tailed test or a two-tailed test ?

(*c*) If the random sample yields an average length of 8.7 inches, would you accept null hypothesis or alternative hypothesis?

32. A random sample of 400 villages was taken from Dhanbad and the average population per village was found to be 527 with a standard deviation of 45. Another random sample of 400 villages was taken from Muzaffarpur where the average population per village was found to be 505 with a standard deviation of 50. Using an appropriate test of significance, state clearly if the difference between the two averages is statistically significant at 5% level.

33. You are given the following information relating to purchase of bulbs from two manufacturers *A* and *B* :

Manufacturer	*No. of Bulbs bought*	*Mean life*	*S.D.*
A	100	2950 hrs.	100 hrs.
B	100	2970 hrs.	90 hrs.

Is there a significant difference in the mean life of two makes of bulbs ?

34. A man buys 200 electric bulbs of each of two well-known makes taken at random from stock for testing purposes. He finds that 'Make *A*' has a mean life of 2,500 hours with a standard deviation of 90 hours and 'Make *B*' has a mean life of 2,650 hours with a standard deviation of 75 hours. Is there a significant difference in the mean life of these two makes at 5% level of significance ?

35. Random samples drawn from two places gave the following data relating to the heights of adult males :

	Place A	*Place B*
Mean height (in inches)	68.50	65.50
Standard deviation (in inches)	2.5	3.0
No. of adult males in sample	1,200	1,500

Test at 5% level, that the mean height is the same for adults in the two places.

36. A stock broker claims that he can predict with 80 per cent accuracy whether a stock's market value will rise or fall during the coming month. As a test, he predicts the outcome of 40 stocks and is correct in 28 of the predictions. Does this evidence support the stock broker's claims ?

37. Two research laboratories have independently produced drugs that provide relief to arthritis patients. The first drug was tested on a group of 100 arthritis patients and produced an average of 8.5 hours of relief with a standard deviation of 2 hours. The second drug was tested on 75 patients, producing an average of 7.8 hours of relief with a standard deviation of 1.5 hours. At a significance level of 1 per cent, does the first drug provide a significantly longer period of relief ?

38. In a simple random sample of 600 men taken from a big city, 450 are found to be smokers. In another simple random sample of 900 men taken from another city, 450 are smokers. Do the data indicate that there is a significant difference in the habit of smoking in the two cities ?

39. An auto company decided to introduce a new six cylinder car whose mean petrol consumption is claimed to be lower than that of the existing auto engine. It was found that the mean petrol consumption for the 50 cars was 14 km per litre with a standard deviation of 3.5 km. per litre. Test for the company at 5% level of significance, whether the claim, the new car petrol consumption is 13.5 km per litre on the average is acceptable.

40. The management of a company claims that the average weekly income of their employees is Rs. 900. The trade union disputes this claim stressing that it is rather less. An independent survey of 150 randomly selected employees estimated the average to be Rs. 856 and the Standard Deviation to be Rs. 364.26. Would you accept the view of the management or the trade union ?

41. Two brands of bulbs are quoted at the same price. A buyer tested a random sample of 100 bulbs of each brand and found the following :

	Mean life (hrs)	*S.D. (hours)*
Brand I	1300	82
Brand II	1248	83

Is there a significant difference in the quality of two brands of bulbs at 5% level of significance ?
[4.45]
(MBA, Delhi Univ., 2006)

42. (*a*) In two large populations, there are 30% and 25% fair coloured people, respectively. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations ? (Given, the tabulated value of test statistics at 5% level of significance is 1.96)
(MBA, IGNOU, 2004)

(*b*) A filling machine at a soft drink factory is designed to fill bottles of 200 ml with a standard deviation of 10 ml. A random sample of 50 filled bottles was taken and the average volume of soft drink was computed to be 198 ml per bottle. Test the hypothesis that the mean volume of soft drink per bottle is not less than 200 ml at 5% level of significance.
(MBA, IGNOU, 2007)

*****

16

Small Sampling Theory

INTRODUCTION

The techniques examined in earlier chapters (13, 14, 15) under the general headings of sampling distributions, estimation of parameters and tests of hypothesis were based on a knowledge of the underlying sampling distribution of the sample statistic for large samples. We have discussed earlier, that if the original population is normally distributed, all sampling distributions of the mean shall be normally distributed regardless of the sample size (central limit theorem). If the original population is normally distributed and the standard deviation of the population is unknown (and therefore, has to be estimated from a sample), the sampling distribution of the mean derived from large samples will also be normally distributed, but if the sample size is small (say 30, or less) then the sample statistic will follow a t-distribution. Problems of estimation and tests of hypothesis for large samples were developed in previous chapters and this chapter extends these concepts for small samples, when the underlying sampling distribution of the mean follows a Student's t-distribution.

The Student's t-distribution obtained by W.S. Gosset was published under the pen name of "Student" in the year 1908. It is reported that Gosset was a statistician for a brewery, and that the management did not want him to publish his scholarly theoretical work under his real name and bring shame to his employer. Consequently, he selected the pen name of Student.

As a matter of fact, procedures of statistical inference for small samples are the same as those presented in preceding two chapters. 'The study of statistical inference with the small samples is called small sampling theory or exact sampling theory'. In this chapter, we shall discuss in detail the "t" and "F" distributions. These two distributions are defined in terms of *number of degrees of freedom.* It is appropriate at this stage to clarify this concept.

Degrees of freedom. The number of degrees of freedom, usually denoted by the Greek symbol ν (read as nu) can be interpreted as the number of useful items of information generated by a sample of given size with respect to the estimation of a given population parameter. Thus, a sample of size 1 generates one piece of useful information if one is estimating the population mean, but none, if one is estimating the population variance. In order to know about the variance, one need at least a sample of size $n \geq 2$. The number of degrees of freedom, in general, is the total number of observations minus the number of independent constraints imposed on the observations.

Suppose the expression $\Sigma X = X_1 + X_2 + X_3$ has four terms. We can arbitrarily assign values to any three of these four values (for example, $15 = X_1 + 2 + 8$) but the value of the fourth is automatically determined (for example, $X_1 = 5$).

In this example, there are 3 degrees of freedom. If n is the number of observations and k is the number of independent constants (the number of constants that have to be estimated from the original data) then $n - k$ is the number of degrees of freedom.

If we consider sample of size n drawn from a normal (or approximately normal) population with mean μ and if for each sample we compute t, using the sample mean $\bar{x}$ and sample standard deviation s, the distribution for t can be obtained. The probability density function of the t-distribution is given by

$$f(t) = \frac{Y_0}{\left(1 + \frac{t^2}{v}\right)^{(v+1)/2}} \qquad -\infty < t < \infty$$

Y_0 is a constant depending on n such that the total area under the curve is one.

$v = n - 1$ is called the number of degrees of freedom.

Properties of *t*-Distribution

(1) The t-distribution ranges from $-\infty$ to ∞ just as does a normal distribution.

(2) The t-distribution like the standard normal distribution is bell-shaped and symmetrical around mean *zero*.

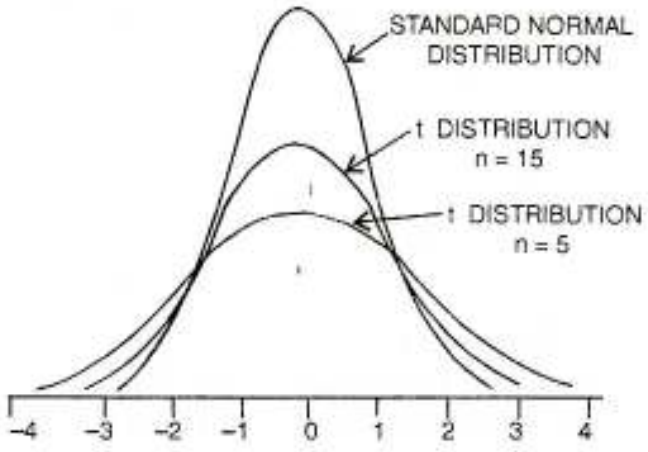
(3) The shapes of the t-distribution changes as the number of degrees of freedom changes. Therefore, for different degrees of freedom, the t-distribution has a family of t-distributions. Hence, the degrees of freedom v is a parameter of the t-distribution.

(4) The variance of the t-distribution is always greater than one and is defined only when $v \geq 3$ and is given as

$$\text{Var}(t) = \left(\frac{v}{v-2}\right)$$

(5) The t-distribution is more of platykurtic (less peaked at the centre and higher in tails) than the normal distribution.

(6) The t-distribution has a greater dispersion than the standard normal distribution. As n gets larger, the t-distribution approaches the normal form. When n is as large as 30, the difference is very small. Relation between the t-distribution and standard normal distribution is shown in the diagram.



Standard Normal Distribution compared with distribution when $n = 5$ and $n = 15$

The t-distribution has different shapes depending on the size of the sample. When the sample is quite small, for example, if n equals five, the height of the t-distribution is shorter than the normal distribution and the tails are wider. As n nears 30, however, the t-distribution approaches the normal distribution in shape.

The *t*-table. The t-table given at the end of the book is the probability integral of t-distribution. It gives over a range of values of v at different levels of significance. By selecting a particular degrees of freedom and level of significance, we determine the tabular value of t. We establish a null hypothesis,

and if our computed t is greater than the tabular t, we reject the null hypothesis ; if our computed t is smaller than the tabular t, we accept the null hypothesis.

Applications of t-distribution. The following are some important applications of the t-distribution :

(1) Test of Hypothesis about the population mean.

(2) Test of Hypothesis about the difference between two means.

(3) Test of hypothesis about the difference between two means with dependent samples.

(4) Test of hypothesis about coefficient of correlation.

(1) **Test of Hypothesis about the Population Mean** (σ unknown and sample size is small). When the population distribution is normal and standard deviation σ is unknown then the "t" statistic is defined as :

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows the Student's t-distribution with $(n-1)$ *d.f.*

where $\bar{x}$ = sample mean

μ = hypothesised population mean

n = sample size

and s is the standard deviation of the sample calculated by the formula :

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$$

The null hypothesis to be tested is whether there is a significant difference between $\bar{x}$ and μ.

If the calculated value of t exceeds the table value of t at a specified level of significance, the null hypothesis is rejected and the difference between $\bar{x}$ and μ is regarded significant. If the calculated value of t is less than the table value, the difference between $\bar{x}$ and μ is not considered to be significant. It may be noted that this test is based on $n-1$ degrees of freedom.

Confidence Interval for the Population Mean. When sampling is from a normally distributed population with unknown σ, the 100 $(1-\alpha)$ per cent confidence interval for the population is given by

$$\bar{x} \pm t_{\alpha/2,\,\nu}\, s/\sqrt{n}$$

Thus,

$$Pr.\,[-t_{\alpha/2,\,\nu} < t < t_{\alpha/2,\,\nu}] = 1-\alpha$$

$$Pr.\left[-t_{\alpha/2,\,\nu} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{\alpha/2,\,\nu}\right] = 1-\alpha$$

Hence, 100 $(1-\alpha)$% confidence interval is given by

$$\bar{x} - t_{\alpha/2,\,\nu}\, s/\sqrt{n} < \mu < \bar{x} + t_{\alpha/2,\,\nu}\, s/\sqrt{n}$$

Illustration 1. Ten oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 15.8 kg and standard deviation is 0.50 kg. Does the sample mean differ significantly from the intended weight of 16 kg ?

(MBA, DU, 1999)

Solution. Let the null hypothesis be that the sample mean weight is not different from the intended weight.

Given that $n = 10,\ \bar{x} = 15.8,\ s = 0.50,\ \mu = 16$

Using the t-test, we have

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{15.8-16}{0.50/\sqrt{10}} = -\frac{0.2}{0.158} = -1.266$$

The table value of t for 9 d.f. at 5% level of significance is 2.26. The computed value of t is smaller than the table value of t. Therefore, the difference is insignificant and the null hypothesis is accepted. Hence the difference between sample mean weight and the intended weight is insignificant.

Illustration 2. Prices of shares (in Rs.) of a company on the different days in a month were found to be : 66, 65, 69, 70, 69, 71, 70, 63, 64 and 68

Test. whether the mean price of the shares in the month is 65. *(MBA, Delhi Univ., 2001)*

Solution. Null hypothesis $H_0 : \mu = 65$, Assuming the population to be normally distributed and the population standard deviation is unknown, the appropriate test statistic to be used is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$\bar{x}$ and s can be computed from the sample values from the following table :

x	$(x-65)$ d	d^2
66	1	1
65	0	0
69	4	16
70	5	25
69	4	16
71	6	36
70	5	25
63	−2	4
64	−1	1
68	3	9
	$\Sigma d = 25$	$\Sigma d^2 = 133$

$$\bar{x} = A + \frac{\Sigma d}{N} = 65 + \frac{25}{10} = 67.5$$

$$s^* = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma x^2}{n-1} - \frac{(\Sigma x)^2}{n(n-1)}} = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}} = \sqrt{\frac{133}{9} - \frac{(25)^2}{10\times 9}}$$

$$= \sqrt{14.78 - 6.94} = \sqrt{7.84} = 2.8$$

Therefore, $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{67.5 - 65}{2.8/\sqrt{10}} = \frac{2.5}{0.89} = 2.81.$$

The table value t for 9 degrees of freedom at 5% level of significance is 2.26. Since the computed value of $t = 2.81$ is greater than the table value, we reject the null hypothesis. Hence, the mean price of the shares in the month is not 65.

(2) Test of Hypothesis about the Difference between Two Means

In testing a hypothesis concerning the difference between the means of two normally distributed populations when the population variances are unknown, the t-test can be used in two types of cases : (*a*) the case in which variances are equal, *i.e.*, $\sigma_1^2 = \sigma_2^2$, (*b*) the case in which variances are not equal, *i.e.*, $\sigma_1^2 \neq \sigma_2^2$.

(*a*) **Case of equal variances.** Let the null hypothesis be that there is no significant difference between the means of the two populations, *i.e.*, $H_0 : \mu_1 = \mu_2$. When the population variances (though unknown) are equal then the appropriate test statistic to be used is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

---

* $\frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{\Sigma(x^2 + \bar{x}^2 - 2x\bar{x})}{n-1} = \frac{\Sigma x^2 - n\bar{x}^2}{n-1} = \frac{\Sigma x^2}{n-1} - \frac{(\Sigma x)^2}{n(n-1)}$

will follow t-distribution with $(n_1 + n_2 - 2)$ d.f., where $\bar{x}_1$ and $\bar{x}_2$ are sample means of sample 1 of size n_1 and sample 2 of size n_2 respectively; μ_1 and μ_2 are the population means, and s is "pooled" estimate of the common population standard deviation obtained by pooling the data from both the samples as given below :

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where $$s_1^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2}{n_1 - 1}; \text{ and } s_2^2 = \frac{\Sigma(x_2 - \bar{x}_2)^2}{n_2 - 1}$$

Therefore, alternatively s can be computed from

$$s = \sqrt{\frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}.$$

If the computed value of t is less than the table value of t at a specified level of significance, the null hypothesis is accepted and the difference between the two means is regarded as insignificant. If the computed value of t is more than the table value of t, the null hypothesis is rejected and the difference between the sample means is regarded as significant.

(*b*) **Case of unequal variances.** When the population variances are not equal, *i.e.*, $\sigma_1^2 \neq \sigma_2^2$, we use the unbiased estimators s_1^2 and s_2^2 to replace σ_1^2 and σ_2^2. In this case, the difficulty arises because the sampling distribution has large variability than the population variability. The statistic :

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

may not strictly follow t-distribution but may be approximated by t-distribution with a modified value for the degrees of freedom given by

$$d.f. = \frac{\left[s_1^2/n_1 + s_2^2/n_2 \right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Illustration 3. Two different types of drugs A and B were tried on certain patients for increasing weight, 5 persons were given drug A and 7 persons were given drug 'B'. The increase in weight (in pounds) is given below :

Drug A :	8	12	13	9	3		
Drug B :	10	8	12	15	6	8	11

Do the two drugs differ significantly with regard to their effect in increasing weight ?

Solution. Null hypothesis $H_0 : \mu_1 = \mu_2$, *i.e.*, there is no significant difference in the efficacy of the two drugs. Applying t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

CALCULATION FOR $\bar{x}_1$, $\bar{x}_2$ AND s

x_1	$(x_1 - \bar{x}_1)$	$(x_1 - \bar{x}_1)^2$	x_2	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
8	–1	1	10	0	0
12	+3	9	8	–2	4
13	+4	16	12	+2	4
9	0	0	15	+5	25
3	–6	36	6	–4	16
			8	–2	4
			11	+1	1
$\Sigma x_1 = 45$	$\Sigma (x_1 - \bar{x}_1) = 0$	$\Sigma (x_1 - \bar{x}_1)^2 = 62$	$\Sigma x_2 = 70$	$\Sigma (x_2 - \bar{x}_2) = 0$	$\Sigma (x_2 - \bar{x}_2)^2 = 54$

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{45}{5} = 9, \quad \bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{70}{7} = 10$$

$$s = \sqrt{\frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{62 + 54}{5 + 7 - 2}} = \sqrt{\frac{116}{10}} = 3.406$$

Therefore,
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{9 - 10}{3.406}\sqrt{\frac{5 \times 7}{5 + 7}} = -\frac{1}{3.406} \times 1.708 = -0.5$$

$v = n_1 + n_2 - 2 = 5 + 7 - 2 = 10$. For $v = 10$, $t_{0.05} = 2.23$.

The calculated value of t is less than the table value. Our null hypothesis is accepted. Hence, we conclude that there is no significant difference in the efficacy of the two drugs in the matter of increasing weight.

Illustration 4. Two salesmen A and B are working in a certain district. From a sample survey conducted by the Head Office, the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen.

	A	B
No. of Sales	10	18
Average sales (in lakh Rs.)	190	205
Standard deviation (in lakh Rs.)	20	25

Solution. Null hypothesis $H_0: \mu_1 = \mu_2$, *i.e.*, there is no significant difference in the average sales between the two salesmen. Applying t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where
$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{9(20)^2 + 7(25)^2}{10 + 18 - 2}} = \sqrt{\frac{3600 + 10625}{26}} = 23.39$$

$$t = \frac{190 - 205}{23.39}\sqrt{\frac{10 \times 18}{10 + 18}} = \frac{15}{23.39} \times 2.54 = 1.63$$

The table value of t at 5% level of significance for 26 *d.f.* is 2.056. The calculated value of t is less than the table value. The hypothesis holds true. Hence, we conclude that there is no significant difference in the average sales between the two salesmen.

Confidence Interval for the Difference between the Two Means

Two samples of sizes n_1 and n_2 are randomly and independently drawn from two normally distributed populations with unknowns but equal variances. The 100 $(1-\alpha)$ per cent confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\, v}\, s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(3) Test of hypothesis about the difference between two means with dependent samples*. In the previous section, we assumed that the two random samples drawn from the two populations were independent. In many practical situations, this may not be true. The samples are *dependent* if they are paired so that each observation in one sample is associated with some particular observation in the second sample. Because of this property, the test we are going to use will be called *paired t-test*. In this test, it is necessary that the observations in the two samples be collected in the form called *matched*

* This is also known as the paired t-test.

pairs. If two samples are dependent, they must have the same number of units. Instead of obtaining two random samples, we can get one random sample of pairs, and the two measurements associated with a pair will be related to each other. This kind of a problem arises in cases such as before and after type experiment or when observations are matched by rise or some other criterion. Suppose that two training methods are to be compared on the basis of average scores by management trainees divided into two equal size classes, one taught by each method. When experimental results are available, we test the null hypothesis that the means associated with the two methods are equal, *i.e.*, $H_0 : \mu_1 = \mu_2$. The appropriate test statistic to be used here is

$$t = \frac{\bar{d}\sqrt{n}}{s}$$

follows t-distribution with $(n-1)$ *d.f.* where $\bar{d}$ = mean of the differences is given by $\bar{d} = \Sigma d/n$, s is the standard deviation of the differences and is given by

$$s = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}} = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}}$$

and n is the number of paired observations in the samples. If the computed value of t (at a specified level of significance for a given number of degrees of freedom) is less than the table value of t, our null hypothesis is accepted, otherwise rejected.

Illustration 5. Ten persons were appointed in an officer cadre in an office. Their performance was noted by giving a test and the marks were recorded out of 100. They were given 4 months training and a test was held and marks were recorded out of 100.

Employee :	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Before training :	80	76	92	60	70	56	74	56	70	56
After training :	84	70	96	80	70	52	84	72	72	50

By applying the t-test, can it be concluded that the employees have benefited by the training?

Solution. Let us take the null hypothesis that the employees have not benefited by the training. Applying t-test :

Employees	*Before* 1*st*	*After* 2*nd*	(1*st* – 2*nd*) *d*	d^2
A	80	84	–4	16
B	76	70	+6	36
C	92	96	–4	16
D	60	80	–20	400
E	70	70	0	0
F	56	52	+4	16
G	74	84	–10	100
H	56	72	–16	256
I	70	72	–2	4
J	56	50	+6	36
$n = 10$			$\Sigma d = -40$	$\Sigma d^2 = 880$

$$t = \frac{\bar{d}\sqrt{n}}{s}, \text{ where } \bar{d} = \frac{\Sigma d}{n} = \frac{-40}{10} = -4$$

$$s = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}}$$

$$= \sqrt{\frac{880 - 10(-4)^2}{9}} = \sqrt{\frac{880-160}{9}} = 8.944$$

$$t = \frac{-4\sqrt{10}}{8.944} = -\frac{4 \times 3.162}{8.944} = -1.414.$$

$v = 10 - 1 = 9$, For $v = 9$, $t_{0.05} = 2.62$.

The calculated value is less than the table value. The null hypothesis holds true. Hence, it can be concluded that the employees have not benefited by the training.

Illustration 6. Ten workers were given a training programme with a view to shorten their assembly time for a certain mechanism. The results of the time and motion studies before and after the training programme are given below :

Worker	:	1	2	3	4	5	6	7	8	9	10
First study (in mnts)	:	15	18	20	17	16	14	21	19	13	22
Second study (in mnts)	:	14	16	21	10	15	18	19	16	14	20

On the basis of this data, can it be concluded that the training programme has shortened the average assembly time?

Solution. Let us take the null hypothesis that the training programme has not helped in reducing the average assembly time. Applying t-test :

$$t = \frac{\bar{d}\sqrt{n}}{s}.$$

CALCULATIONS FOR $\bar{d}$ AND s

Worker	*1st study*	*2nd study*	*(1st–2nd)* d	d^2
1	15	14	+1	1
2	18	16	+2	4
3	20	21	–1	1
4	17	10	+7	49
5	16	15	+1	1
6	14	18	–4	16
7	21	19	+2	4
8	19	16	+3	9
9	13	14	–1	1
10	22	20	+2	4
			$\Sigma d = 12$	$\Sigma d^2 = 90$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{12}{10} = 1.2.$$

$$s = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}} = \sqrt{\frac{90}{9} - \frac{(12)^2}{10 \times 9}} = \sqrt{10 - 1.6} = \sqrt{8.4} = 2.898$$

$$t = \frac{\bar{d}\sqrt{n}}{s} = \frac{1.2\sqrt{10}}{2.898} = \frac{1.2 \times 3.162}{2.898} = 1.309.$$

For $v = 9$, the table value of t at 5% level of significance is 2.262. Since the computed value of t is less than the table value, the null hypothesis is accepted. Hence, the training programme has not shortened the average assembly time.

Confidence Interval for the Mean of the Difference. When the population for the mean of differences is normally distributed with unknown variance for dependent samples, a 100 $(1 - \alpha)$ per cent confidence interval is given by

$$\bar{d} \pm t_{\alpha\, 2,v}\ s/\sqrt{n}\ .$$

(4) Test of Hypothesis about Coefficient of Correlation

Case 1 : Testing the hypothesis when the population coefficient of correlation equals zero, *i.e.*, $H_0 : \rho = 0$.

Here the null hypothesis is that there is no correlation in the population, *i.e.*, $H_0 : \rho = 0$. The population coefficient of correlation ρ measures the degree of relationship between the variables. When $\rho = 0$, there is no statistical relationship between the variables. In order to test this hypothesis, it is necessary to know the sample coefficient of correlation r (which is the best estimate of ρ). The appropriate test statistic to be used here is given by :

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

which follows t-distribution with $n-2$ degrees of freedom.

If the computed value of t is greater than the table value of t, the null hypothesis is rejected which indicates that sample data provides sufficient evidence to indicate that $\rho \neq 0$. Hence, it can be concluded that there is a linear relationship between the variables.

Illustration 7. In a study of the relationship between expenditure (X) and annual sales volume (Y), a sample of 10 firms yielded the coefficient of correlation $r = 0.93$. Can we conclude on the basis of this data that X and Y are linearly related ?

Solution. The null hypothesis is $H_0 : \rho = 0$, *i.e.*, there is no relationship between two variables. Using the t-test

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2} = \frac{0.93}{\sqrt{1-(0.93)^2}} \sqrt{10-2}$$

$$= \frac{0.93}{\sqrt{0.14}} \times \sqrt{8} = \frac{0.93 \times 2.828}{0.374} = 7.03.$$

The degrees of freedom or $v = n - 2 = 10 - 2 = 8$.

The table value of t at 5% level of significance for 8 *d.f.* is 2.306. Since the computed value is much greater than the table value, the null hypothesis is rejected. Hence, it may be concluded that X and Y are linearly related.

Case 2 : Testing the hypothesis when the population coefficient of correlation equals some other value than zero, *i.e.*, $H_0 : \rho = \rho_0$.

In this case when $\rho \neq 0$, the test based on t-distribution will not be appropriate. In testing the hypothesis, the use of Fisher's z-transformation will be applicable. Here, r is transformed into z by

$$z = \frac{1}{2} \log_e \frac{1+r}{1-r}$$

Here, $\log_e$ is a *natural* logarithm. Common logarithms may be shifted to natural logarithms by multiplying by the factor 2.3026, *i.e.*,

$$\log_e X = 2.3026 \log_{10} X$$

where X is a positive integer.

Since $\frac{1}{2}(2.306) = 1.1513$, the transformation formula may be used as :

$$z = 1.1513 \log_{10} \frac{1+r}{1-r}$$

Now, it can be shown that Z is normally distributed with mean

$$z_\rho = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho}$$

and standard deviation

$$\sigma_z = \frac{1}{\sqrt{n-3}}$$

Therefore, to test the null hypothesis that $\rho = \rho_0$, the test statistic would be :

$$z = \frac{z - z_\rho}{\sigma_z}$$

which follows approximately the standard normal distribution. This test is more appropriate if sample size is large. The approximation is reasonably good if the sample size is at least 10.

Case 3 : Testing the hypothesis for the difference between two independent correlation coefficients.

To test the hypothesis of two correlation coefficients derived from two separate samples, we have to compare the difference of the two corresponding values of z with the standard error of that difference. In this case, the formula used will be

$$z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

where $$z_1 = \frac{1}{2} \log_e \frac{1 + r_1}{1 - r_2} = 1.1513 \log_{10} \frac{1 + r_1}{1 - r_1}$$

and $$z_2 = \frac{1}{2} \log_e \frac{1 + r_2}{1 - r_2} = 1.1513 \log_{10} \frac{1 + r_2}{1 - r_2}.$$

If the absolute value of this statistic is greater than 1.96, the difference will be significant at 5% level.

Illustration 8. The following data give sample size :

Sample Size	Value of r
5	0.87
12	0.56

Test the significance of the difference between two values using Fisher's z-transformation.

Solution. Let the null hypothesis be that the experiment provides no evidence that the samples are drawn from the same population. Applying z-test :

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

$$z_1 = \frac{1}{2} \log_e \frac{1 + r_1}{1 - r_1} = 1.1513 \log_{10} \frac{1 + r_1}{1 - r_1}$$

$$= 1.1513 \log_{10} \frac{1 + 0.87}{1 - 0.87} = 1.1513 \log_{10} \frac{1.87}{0.13}$$

$$= 1.1513 \log_{10} 14.385 = 1.1513 \times 1.579 = 1.82$$

$$z_2 = \frac{1}{2} \log_e \frac{1 + r_2}{1 - r_2} = 1.1513 \log_{10} \frac{1 + r_2}{1 - r_2}$$

$$= 1.1513 \log_{10} \frac{1 + 0.56}{1 - 0.56} = 1.1513 \log_{10} = \frac{1.56}{0.44}$$

$$= 1.1513 \log 3.545 = 1.1513 \times 0.5495 = 0.63$$

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{1.82 - 0.63}{\sqrt{\frac{1}{5 - 3} + \frac{1}{12 - 3}}}$$

$$= \frac{1.19}{\sqrt{0.5 + 0.1}} = \frac{1.19}{\sqrt{0.6}} = \frac{1.19}{0.77} = 1.54.$$

Since the computed value of z is less than the table value of z at 5% level of significance, therefore, the null hypothesis is accepted. Hence, the experiment provides no evidence against the hypothesis that the samples are drawn from the same population.

Illustration 9. Suppose we want to test whether $r = 0.884$ for a pair of 20 observations is significantly different from a hypothesised value $\rho = 0.92$.

Solution. We transform r into z by :

$$z = 1.5113 \log_{10} \frac{1 + 0.884}{1 - 0.884} = 1.3938.$$

The distribution of z is approximately normal around the hypothesised value

$\rho_0 = 0.92 = z\rho_0$, where

$$z\rho_0 = 1.1513 \log_{10} \frac{1 + 0.92}{1 - 0.92} = 1.5890.$$

The distribution of z has a standard deviation

$$\sigma_z = \frac{1}{\sqrt{20-3}} = \frac{1}{\sqrt{17}} = \frac{1}{4.123} = 0.2425.$$

Therefore, the statistic is

$$z = \frac{1.3938 - 1.5890}{0.2425} = -0.80.$$

From the table of areas for the normal curve, we find that in about 20 per cent times we may expect a difference as large or larger than this. This hypothesis that $r = 0.884$ can be rejected at a low level of confidence.

The *F*-Distribution

The *F*-distribution is named in honour of R.A. Fisher who *first studied it* in 1924. This distribution is usually defined in terms of the ratio of the variances of two normally distributed populations. The quantity

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

is distributed as *F*-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom,

where $s_1^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2}{n_1 - 1}$ is the unbiased estimator of σ_1^2 and $s_2^2 = \frac{\Sigma(x_2 - \bar{x}_2)^2}{n_2 - 1}$ is the unbiased estimator of σ_2^2.

If $\sigma_1^2 = \sigma_2^2$, then the statistic

$$F = \frac{s_1^2}{s_2^2}$$

follows *F*-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

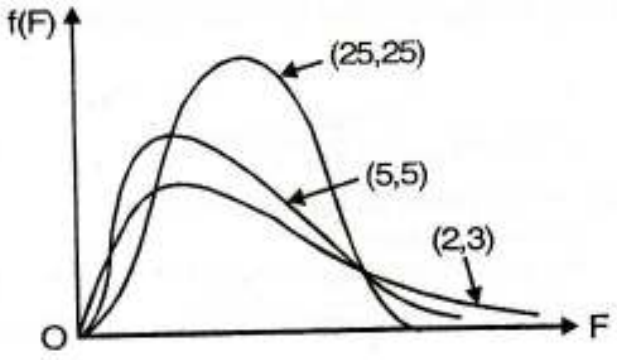
The *F*-distribution sometimes is also called *Variance Ratio* distribution which can be seen from the definition. The *F*-distribution depends on the degrees of freedom v_1 for the numerator and v_2 for the denominator. Therefore, the parameters for *F*-distribution are v_1 and v_2. For different values of v_1 and v_2 we shall have different distributions.

The probability density function of *F*-distribution is given by

$$f(F) = Y_0 \frac{F_{v_1/2-1}}{\left[1 + \frac{v_1}{v_2}\right](v_1 + v_2)/2} \qquad 0 \le F \le \infty$$

where Y_0 is a constant depending on the values v_1 and v_2 such that the area under the curve is unity. A typical *F*-distribution is given as below :

Some of the important properties of *F*-distribution are given below :



(1) The F-distribution is positively skewed and its skewness decreases with increase in v_1 and v_2.

(2) The value of F must always be positive or zero since variances are squares and can never assume negative values. Its value will always lie between 0 and ∞.

(3) The mean and variance of the F-distribution are

$$\text{Mean} = \frac{v_2}{-v_2 - 2}, \text{ for } v_2 > 2$$

$$\text{Variance} = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}, \text{ for } v_2 > 4.$$

(4) The shape of the F-distribution depends upon the number of degrees of freedom.

(5) The areas in the left-hand side of the distribution can be found by taking the reciprocal of F values corresponding to the right-hand side, when the number of degrees of freedom in the numerator and in the denominator are interchanged. This is also known as *reciprocal property* and *can be expressed as*

$$F_{1-\alpha, v_1, v_2} = \frac{1}{F_{\alpha, v_2, v_1}}$$

where the symbols have their usual meanings. This property is of great help when we want to know the lower tail F values from corresponding upper tail F values which are given in the Appendix.

Testing of Hypothesis for Equality of two Variances

The test of equality of two population variances is based on the variances in two independently selected random samples drawn from two normal populations. Under the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$.

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \text{ reduces to } F = \frac{s_1^2}{s_2^2}.$$

which follows F-distribution with v_1 and v_2 degrees of freedom. It is convenient to place larger sample variance in the numerator for computational purpose. If we do so, the ratio of the sample variance will be equal to or greater than one.

If the computed value of F exceeds the table value of F, we reject the null hypothesis, *i.e.*, the alternate hypothesis is accepted.

Illustration 10. Two sources of raw materials are under consideration by a company. Both sources seem to have similar characteristics but the company is not sure about their respective uniformity. A sample of 10 lots from source A yields a variance of 225 and a sample of 11 lots from source B yields a variance of 200. Is it likely that the variance of source A is significantly greater than the variance of source B at $\alpha = 0.01$?

Solution. Null hypothesis is $H_0 : \sigma_1^2 = \sigma_2^2$, *i.e.*, the variances of source A and that of source B are same. The F statistic to be used here is

$$F = \frac{s_1^2}{s_2^2}$$

where $s_1^2 = 225$, and $s_2^2 = 200$

$$F = \frac{225}{200} = 1.1$$

The table value of F for $v_1 = 9$ and $v_2 = 10$ at 1% level of significance is 4.49. Since the computed value of F is smaller than the table value of F, the null hypothesis is accepted. Hence, the population variances of the two populations are same.

Confidence Interval for the Ratio of Two Variances

A 100 $(1-\alpha)$ per cent confidence interval for the ratio of the variances of two normally distributed populations is given by

$$\frac{s_1^2/s_2^2}{F_{(1-\alpha/2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2/s_2^2}{F_{\alpha/2}}$$

where the symbols have their usual meanings.

MISCELLANEOUS ILLUSTRATIONS

Illustration 11. The nine items of a sample had the following values :

45, 47, 50, 52, 48, 47, 49, 53, 50

The mean is 49 and the sum of squares of deviations taken from mean is 52. Can this sample be regarded as taken from the population having 47 as mean ? Also obtain 95% and 99% confidence limits of the population mean. *(MBA, Delhi Univ., 2002)*

Solution. The null hypothesis is $H_0 : \mu = 47$, *i.e.*, the population mean is 47, we are given that

$\bar{x} = 49$, $\Sigma (x - \bar{x})^2 = 52$, and $\mu = 47$, $n=9$. Applying t-test

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} = \sqrt{\frac{52}{8}} = 2.55$$

Substituting the values, we have

$$t = \frac{49 - 47}{2.55/3} = 2.35$$

The table value of t for 8 *d.f.* at 5% level of significance is 2.31. Since the computed value is slightly greater than the table value, the null hypothesis is rejected. Hence, the samples are not drawn from the population having 47 as mean.

95% confidence interval of the population mean is given by :

$$\bar{x} \pm t_{0.05}\ s/\sqrt{n}$$

$$= 49 \pm \frac{2.31 \times 2.55}{3} = 49 \pm 1.96 = 47.04 \text{ to } 50.96$$

99% confidence interval of the population mean is given by

$$\bar{x} \pm t_{0.01}\ s/\sqrt{n}$$

$$= 49 \pm \frac{3.36 \times 2.55}{3} = 49 \pm 2.86 = 46.14 \text{ to } 51.86.$$

Illustration 12. A company is interested in knowing if there is a significant 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 standard deviations are about the same.

	First division	*Second division*
Sample size	12	10
Average weekly salary of foremen (Rs.)	1050	980
Standard deviation of salaries (Rs.)	68	74

Solution. Let us take the null hypothesis that the average salary received by foremen in the two divisions does not differ significantly. Applying t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(12-1)68^2 + (10-1)74^2}{12+10-2}} = \sqrt{\frac{50864 + 49284}{20}} = 70.76$$

$$t = \frac{1050 - 980}{70.76} = \sqrt{\frac{12 \times 10}{12+10}} = \frac{70}{70.76} \times 2.34 = 2.31$$

$$v = n_1 + n_2 - 2 = 12 + 10 - 2 = 20.$$

For $\nu = 20$, $t_{0.05} = 2.086$. The calculated value of t is greater than the table value. The null hypothesis does not hold good. Hence, there is significant difference in the salary received by foremen in the two divisions.

Illustration 3. Ten accountants were given intensive coaching and four tests were conducted in a month. The scores of tests 1 and 4 are given below:

S. No. of Accountants	:	1	2	3	4	5	6	7	8	9	10
Marks in 1st test	:	50	42	51	42	60	41	70	55	62	38
Marks in 4th test	:	62	40	61	52	68	51	64	63	72	50

Does the score from test 1 to test 4 show an improvement ? Test at 5% level of significance.

Solution. Let us take the null hypothesis that there is no improvement from test 1 to 4. Applying t-test:

$$t = \frac{\bar{d}\sqrt{n}}{s}.$$

CALCULATION FOR $\bar{d}$ AND s

S. No.	*1st test*	*4th test*	*(4th–1st)* d	d^2
1	50	62	+12	144
2	42	40	–2	4
3	51	61	+10	100
4	42	52	+10	100
5	60	68	+8	64
6	41	51	+10	100
7	70	64	–6	36
8	55	63	+8	64
9	72	62	+10	100
10	38	50	+12	144
			$\Sigma d = 72$	$\Sigma d^2 = 856$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{72}{10} = 7.2$$

$$s = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{856 - 10(7.2)^2}{9}} = \sqrt{\frac{856 - 518.4}{9}} = 6.125$$

$$t = \frac{7.2\sqrt{10}}{6.125} = \frac{7.2 \times 3.162}{6.125} = 3.72$$

For $\nu = 9$, $t_{0.05} = 1.83$

The calculated value of t is greater than table value. The null hypothesis is rejected. Hence, the scores from test 1 to test 4 show an improvement.

Illustration 14. The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population?

(MBA, Delhi Univ., 2006)

Solution. Let us take the null hypothesis that the samples are drawn from the same normal population. Applying t-test, *i.e.*,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s = \sqrt{\frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{26.94 + 18.73}{9 + 7 - 2}} = \sqrt{\frac{45.67}{14}} = 1.81$$

$$t = \frac{196.42 - 198.82}{1.81}\sqrt{\frac{9 \times 7}{9 + 7}} = \frac{2.40 \times 1.984}{1.81} = \frac{4.76}{1.81} = 2.63$$

For $\nu = 14$, $t_{0.05} = 2.145$. Since the calculated value of t is greater than the table value, we reject the null hypothesis. Hence, the difference between the two means is significant. Therefore, the samples cannot be said to have been drawn from the same normal population.

Illustration 15. Strength tests carried out on samples of two yarns spun to the same count gave the following results :

	Sample size	*Sample mean*	*Sample variance*
Yarn *A*	4	52	42
Yarn *B*	9	42	56

The strengths are expressed in pounds. Is the difference in mean strengths significant of real difference in the mean strengths of the sources from which the samples are drawn ? *(MBA. Delhi Univ., 2004, 2007)*

Solution. Let us take the null hypothesis that there is no significant difference in the mean strengths of the two types of yarns. Applying *t*-test,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where,

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{3\times 42 + 8\times 56}{4+9-2}} = \sqrt{\frac{574}{11}} = 7.224$$

$$t = \frac{52-42}{7.224}\sqrt{\frac{4\times 9}{4+9}} = \frac{10\times 1.664}{7.224} = 2.303$$

$$\nu = n_1 + n_2 - 2 = 4 + 9 - 2 = 11. \text{ For } \nu = 11,\ t_{0.05} = 1.796$$

The calculated value of *t* is more than the table value of *t*. The null hypothesis is rejected. The difference in the mean strengths of the two types of yarn is significant.

Illustration 16. To verify whether a course in accounting improved performance, a similar test was given to 12 participants both before and after the course. The original marks recorded in alphabetical order of the participants were 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53, and 72. After the course, the marks were in the same order : 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60, and 78. Test whether the course was useful ?

Solution. Let us take the null hypothesis that the course has not improved the performance of the participants. Applying *t*-test,

$$t = \frac{\bar{d}\sqrt{n}}{s}$$

CALCULATION FOR $\bar{d}$ AND s

Before	*After*	*(2nd–1st)* d	d^2
44	53	+9	81
40	38	–2	4
61	69	+8	64
52	57	+5	25
32	46	+14	196
44	39	–5	25
70	73	+3	9
41	48	+7	49
67	73	+6	36
72	74	+2	4
53	60	+7	49
72	78	+6	36
		$\Sigma d = 60$	$\Sigma d^2 = 578$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{60}{12} = 5$$

$$s = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{578 - 12(5)^2}{12-1}} = \sqrt{\frac{578-300}{11}} = 5.027$$

$$t = \frac{5\sqrt{12}}{5.027} = \frac{17.32}{5.027} = 3.45$$

For $\nu = 11$, $t_{0.05} = 2.201$. The calculated value of *t* is more than the table value of *t*. The null hypothesis is rejected. Hence, course has improved the performance of the participants.

Illustration 17. Samples of two different types of bulbs were tested for length of life, and the following data were obtained :

	Type I	*Type II*
Sample Size	8	7
Sample Mean	1234 hrs.	1136 hrs.
Sample S.D.	36 hrs.	40 hrs.

Is the difference in mean life of two types of bulbs significant ? *(MBA, Delhi Univ., 2003)*

Solution. Let us take the null hypothesis that there is no significant difference in the mean life of the two types of bulbs. Applying t-test,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{(8-1)\,36^2 + (7-1)\,40^2}{8+7-2}} = \sqrt{\frac{9072+9600}{13}} = 37.9$$

$$t = \frac{1234-1136}{37.9}\sqrt{\frac{8 \times 7}{8+7}} = \frac{98}{37.9} \times 1.932 = 5$$

For $\nu = 13$, $t_{0.05} = 2.16$. The calculated value of t is greater than the table value. The null hypothesis is rejected. Hence, there is a significant difference in the mean life of two types of bulbs.

Illustration 18. Two random samples drawn from 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 population and test whether two populations have the same variance.

Solution. Let us take the null hypothesis that two population have the same variance. Applying F-test :

$$F = \frac{s_1^2}{s_2^2}$$

Sample I			*Sample II*		
x_1	$(x_1 - \bar{x}_1)$	$(x_1 - \bar{x}_1)^2$	x_2	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
20	–2	4	27	–8	64
16	–6	36	33	–2	4
26	+4	16	42	+7	49
27	+5	25	35	0	0
23	+1	1	32	–3	9
22	0	0	34	–1	1
18	–4	16	38	+3	9
24	+2	4	28	–7	49
25	+3	9	41	+6	36
19	–3	9	43	+8	64
			30	–5	25
			37	+2	4
$\Sigma x_1 = 220$		$\Sigma (x_1 - \bar{x}_1)^2 = 120$	$\Sigma x_2 = 420$		$\Sigma (x_2 - \bar{x}_2)^2 = 314$

$$s_1^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2}{n_1 - 1} = \frac{120}{9} = 13.333 ;$$

$$s_2^2 = \frac{\Sigma(x_2 - \bar{x}_2)^2}{n_2 - 1} = \frac{314}{11} = 28.545$$

$$F = \frac{s_1^2}{s_2^2} = \frac{13.333}{28.545} = 0.467$$

Since numerator is greater than denominator, therefore,

$$F = \frac{28.545}{13.33} = 2.14$$

The critical value of F for $v_1 = 9$ and $v_2 = 11$ at 5% level is 4.63. Since the calculated value of F is less than the table value, therefore, there is no reason to reject the null hypothesis. Hence, it may be concluded that the two populations have the same variance.

Illustration 19. A drug manufacturer has installed a machine which fills automatically 5 gms of drug in each phial. A random sample of phials was taken and it was found to contain 5.02 gms on an average in a phial. The S.D. of the sample was 0.002 gms. Test at 5% level of significance, if the adjustment in the machine is in order. *(MBA, DU, 1999)*

Solution. Let us take the null hypothesis that there is no significant difference between $\bar{x}$ and μ. Applying t-test,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$\bar{x} = 5.02, \mu = 5, s = 0.002, n = 10$$

$$t = \frac{5.02 - 5}{0.002} \times \sqrt{10} = \frac{0.02}{0.002} \times 3.162 = 31.62$$

For $v = 9$, $t_{0.05} = 1.833$. The calculated value is much higher than the table value. The null hypothesis is rejected. Hence, the adjustment in the machine is not in order.

Illustration 20. A random sample of 12 families in one city showed an average weekly food expenditure of Rs. 1380 with a standard deviation of Rs. 100 and a random sample of 15 families in another city showed an average monthly food expenditure of Rs. 1320 with a standard deviation of Rs. 120. Test, whether the difference between the two means is significant at a level of significance of 0.01.

Solution. Let us take the null hypothesis that there is no significant difference in the mean expenditure of the families in the two cities. Applying t-test,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(12-1)100^2 + (15-1)120^2}{12+15-2}}$$

$$= \sqrt{\frac{110000 + 201600}{25}} = 111.64$$

$$t = \frac{1380 - 1320}{111.64} \sqrt{\frac{12 \times 15}{12+15}} = \frac{60 \times 2.582}{111.64} = 1.39$$

$$v = 12 + 15 - 2 = 25$$

For $v = 25$, $t_{0.01} = 2.485$.

The calculated value t is less than the table value. The null hypothesis is accepted. Hence, the difference between the two means is not significant.

Illustration 21. Eight students were given a test in statistics, and after one month's coaching, they were given another test of the similar nature. The following table gives the increase in their marks in the second test over the first :

Roll No.	:	1	2	3	4	5	6	7	8
Increase in marks	:	4	–2	6	–8	12	5	–7	2

Do the marks indicate that the students have gained from the coaching ?

Solution. Let us take the null hypothesis that the students have not gained from the coaching. Applying the t-test :

$$t = \frac{\bar{d}\sqrt{n}}{s}$$

d	d^2
+4	16
−2	4
+6	36
−8	64
+12	144
+5	25
−7	49
+2	4
$\Sigma d = 12$	$\Sigma d^2 = 342$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{12}{8} = 1.5$$

$$s = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{342 - 8(1.5)^2}{8-1}} = \sqrt{\frac{342-18}{7}} = 6.8$$

$$t = \frac{1.5\sqrt{8}}{6.8} = 0.624$$

For $\nu = 7$, $t_{0.05} = 1.895$.

The calculated value of t is less than the table value. The null hypothesis is accepted. Hence, the students have not gained from the coaching.

Illustration 22. You are given the following data about the life of two brands of bulbs :

	Mean life	*Standard deviation*	*Sample size*
Brand A	2,000 hrs	250 hrs	12
Brand B	2,230 hrs	300 hrs	15

Do you think there is a significant difference in the quality of the two brands of bulbs ? *(MBA, DU, 2002)*

Solution :

Given : $\bar{x}_1 = 2000$, $s_1 = 250$, $n_1 = 12$

$\bar{x}_2 = 2230$, $s_2 = 300$, $n_2 = 15$

Let the null hypothesis be that there is no significant difference in the quality of the two brands of bulbs.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where,

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{11(62500) + 14(90000)}{12+15-2}}$$

$$= \sqrt{\frac{1947500}{25}} = \sqrt{77900} = 279.11$$

$$t = \frac{2000-2230}{279.11}\sqrt{\frac{12\times 15}{12+15}} = \frac{-230}{279.11}\sqrt{\frac{180}{27}}$$

$$= \frac{-230(2.58)}{279.11} = -\frac{593.4}{279.11} = -2.126$$

The table value of t for 25 *d.f.* at 5% level of significance is 1.708. Since computed value is greater than the table value, therefore, we reject the null hypothesis. Hence, the quality of the two brands of bulbs differ significantly.

Illustration 23. Samples of final examination scores for two statistics classes with different instructors provided the following results :

	Sample Size	*Mean*	*Standard Deviation*
Instructor A	12	72	8
Instructor B	15	78	10

Test, whether these data are sufficient to conclude that the mean scores for the two classes differ. *(MBA, D.U., 2003)*

Solution. Let us take the hypothesis that there is no significant difference in the mean scores because of different instructors. Applying *t*-test of difference of means :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(12-1)\,8^2 + (15-1)10^2}{12 + 15 - 2}}$$

$$= \sqrt{\frac{704 + 1400}{25}} = \sqrt{\frac{2104}{25}} = 9.14$$

$$t = \frac{72 - 18}{9.17}\sqrt{\frac{12 \times 15}{12 + 15}} = \frac{-6}{9.17} \times 2.58 = -1.69$$

For $\nu = 25, t_{0.05} = 2.06$.

The calculated value of *t* is less than the table value. Hence, there is no significant difference in the mean scores of different instructors.

Illustration 24. The average middle class family spends Rs. 9,000 per month. A random sample of 25 families in a city showed a sample mean monthly expenditure of Rs. 8,450 with a standard deviation of Rs. 1,450. Test $H_0 : \mu =$ Rs. 9,000 and $H_a \neq$ Rs. 9,000 with $\alpha = 0.05$. Use Two Tailed test.

(*i*) What are the critical values of the test statistic, and what is the rejection region ?

(*ii*) Compute the value of the test statistic.

(*iii*) What is your conclusion ?

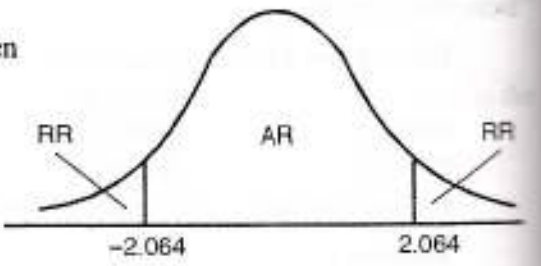
(MBA, Delhi Univ., 2007)

Solution. Given $\mu = 9000, n = 25, \bar{x} = 8450, s = 1450$

Let us take the null hypothesis that there is no significant difference between sample mean monthly expenditure and population monthly expenditure, *i.e.*,

$H_0 : \mu =$ Rs. 9000, $H_a : \mu \neq$ Rs. 9000, Using *t*-test.

(*i*) Critical values of *t* for 24 *d.f.* at 5% level of significance are ± 2.064.



(*ii*) $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} = \dfrac{8450 - 9000}{1450/\sqrt{25}} = -\dfrac{550 \times 5}{1450} = -1.897$

(*iii*) Since the computed value of $t = -1.897$ is less than the table value of $t = -2.064$, therefore, it lies in the acceptance region. Hence, there is no significant difference between sample mean monthly expenditure and population mean monthly expenditure. Therefore, the samples have been drawn from the given population.

Illustration 25. A market research firm used a sample of individuals to rate the purchase potential of a particular product before and after the individuals saw a new television commercial about the product. The purchase potential ratings were based on 0 to 10 scale, with higher values indicating a higher purcahse potential. Test the hypothesis that the commercial improved the mean purchase potential rating. Use level of significance 5% and comment on the value of the commercial.

Individual	:	1	2	3	4	5	6	7	8
Purchase rating (After)	:	6	6	7	4	3	9	7	6
Purchase rating (Before)	:	5	4	7	3	5	8	5	6

(MBA, Delhi Univ., 2009)

Solution. Let us take the hypothesis that the new television commercial has not improved the mean purchase potential rating. Applying the *t*-test :

Individual	*Purchase rating before 1st*	*Purchase rating after 2nd*	*(2nd – 1st) d*	d^2
1	5	6	1	1
2	4	6	2	4
3	7	7	0	0
4	3	4	1	1
5	5	3	–2	4
6	8	9	1	1
7	5	7	2	4
8	6	6	0	0
			$\Sigma d = 5$	$\Sigma d^2 = 15$

$$t = \frac{\bar{d}\sqrt{n}}{s}$$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{5}{8} = 0.625$$

$$s = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}}$$

$$= \sqrt{\frac{15 - 8(0.625)^2}{8-1}} = \sqrt{\frac{15 - 3.125}{7}} = \sqrt{\frac{11.875}{7}} = 1.302$$

$$t = \frac{0.625\sqrt{8}}{1.302} = \frac{0.625 \times 2.828}{1.302} = \frac{1.7675}{1.302} = 1.358$$

$$\nu = 8 - 1 = 7, \quad \text{For } \nu = 7,\ t_{0.05} = 1.895$$

Since the calculated value of t (1.358) is less than the table value (1.895), the null hypothesis is accepted. Hence the new television commercial has not improved the mean purchase potential rating.

Illustration 26. The variance in production process is an important measure of the quality process. A large variance of ten signals an opportunity for improvement in the process by finding ways to reduce the process variance. Conduct a statistical test to determine whether there is a significant difference between the variances in the bag weights for the two machines. Use a 10% level of significance. What is your conclusion? Which machine, if either, provides the greater opportunity for quality improvements ?

	No. of observation	*Mean*	*Standard deviation*
Machine 1 :	25	5.9	2.0
Machine 2 :	22	6.3	1.9

(MBA, Delhi Univ., 2009)

Solution. Let us take the hypothesis that there is no significant difference in the two machines for providing opportunity in quality improvement. Applying the t-test of difference of means :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(25-1)4 + (22-1)3.61}{25 + 22 - 2}} = \sqrt{\frac{96 + 75.81}{45}} = \sqrt{\frac{171.81}{45}} = 1.954$$

$$\bar{X}_1 = 5.9,\ \bar{X}_2 = 6.3,\ n_1 = 25,\ n_2 = 22,\ s = 1.954$$

$$t = \frac{5.9 - 6.3}{1.954}\sqrt{\frac{25 \times 22}{25 + 22}}$$

$$= \frac{0.4}{1.954} \times 3.421 = 0.7$$

$$\nu = 25 + 22 - 2 = 45$$

For $\nu = 45,\ t_{0.10} = 1.301$

Since the calculated value of t is less than the table value, the hypothesis is accepted. Hence there is no significant difference in the two machines for providing opportunity in quality improvement.

PROBLEMS

1-A : Answer the following questions, each question carries **one** mark:

(*i*) What is a t-distribution ?

(*ii*) Write the formula for difference of two means in case of small sample tests.

(*iii*) Define t-test. *(M. Com., Madurai-Kamaraj Univ., 2002)*

(*iv*) Give two important properties of t-distribution.

(*v*) Give at least two important applications of t-distribution.

(*vi*) What do you understand by degrees of freedom ?

(*vii*) What is paired t-test? Give its formula.

(*viii*) What is F-distribution ?

(*ix*) Give two important properties of F-distribution.

(*x*) Give any important application of F-distribution.

1-B : Answer the following questions, each question carries **four** marks:

(*i*) Explain the difference between the means of two samples by using *t*-distribution.

(*ii*) Explain *t*-test distribution and note its properties.

(*iii*) In what ways small sampling theory differs from large sampling theory.

(*iv*) What is the procedure involved in testing hypothesis of a coefficient of correlation.

(*v*) What are the assumptions involved in using the *F*-test for testing the equality of two sample variances?

2. How does small sampling theory differ from large sampling theory ?

3. (*a*) What is *t*-distribution ? Give its important properties.

(*b*) What is Students '*t*' distribution? Point out its usefulness.

4. (*a*) Give some important applications of the *t*-test and explain how it helps in arriving at business decisions.

(*b*) How can "*t*" test be applied for testing the significance of the difference between two sample means ?

5. Discuss the *F*-test for testing the equality of two sample variances. State clearly assumptions involved.

6. 12 persons were appointed in clerical position in an office. Their performance was noted by giving a test and the marks recorded out of 10. They were given 3 month's training and again they were given a test and marks recorded out of 12.

Employees	:	A	B	C	D	E	F	G	H	I	J	K	L
Before training	:	4	5	3	7	8	6	5	9	10	6	4	3
After training	:	5	4	6	8	7	5	9	9	10	6	5	4

Can it be concluded that the training has improved the performance of the employees ?

7. A random sample of 25 pairs of observations from a normal population gives a correlation coefficient of 0.46. Is it likely that the variables in the population uncorrelated ?

8. 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.

9. Two laboratories *A* and *B* carry out independent estimates of fat content in ice-cream made by a firm. A sample is taken from each batch, halved and the separate halves sent to the two laboratories. The fat content obtained by the laboratories is recorded below :

Batch No. :	1	2	3	4	5	6	7	8	9	10
Lab. *A* :	3	5	7	3	8	6	9	4	7	8
Lab. *B* :	9	8	8	4	7	7	9	6	6	6

Is there a significant difference between the mean fat content obtained by the two laboratories *A* and *B* ?

10. An automobile tyre manufacturer claims that the average life of certain grade of tyre is greater than 25,000 km when used under normal driving conditions on a car of a certain weight. A random sample of 15 tyres was tested, and a mean and standard deviation of 27,000 and 5,000 kms respectively, were computed. Can we conclude that the manufacturer's product is as good as claimed ? *(M. Com., Madurai-Kamaraj Univ., 2007)*

[$t = 1.55$]

11. The quality control department of a food processing firm specifies that the mean net weight per package of a certain food must be 20 gms. Experience has shown that the weights are approximately normally distributed with a standard deviation of 15 gms. If a random sample of 15 packages yields a mean weight of 19.5 gms, is this sufficient evidence to indicate that the true mean weight of the package has decreased ?

[$t = 1.29$]

12. Two working designs are under consideration for adoption in a plant. A time and motion study shows that 12 workers using design *A* have mean assembly time of 300 seconds with a standard deviation of 12 seconds and that 15 workers using design *B* have a mean assembly time of 335 seconds with a standard deviation of 15 seconds. Is the difference in the mean assembly time between the two working designs significant at 1% level of significance ?

[$t = 23.52$]

13. The mean life of a sample of 10 electric light bulbs was found to be 1,456 hours with standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1,280 hours with standard deviation of 398 hours. Is there a significant difference between the means of the two batches ? *(MBA, Kumaun Univ., 2009)*

[$t = 1.085$]

14. The variability in the tensile strength of two types of steel wire is to be compared. Given a sample of 14 observations of type *A* wire yielding a variance of 31.5, and a sample of 15 observations of type *B* wire yielding a variance of 29.3. Test the hypothesis that the two populations have equal variances.

15. In an *F* ratio with $\nu_1 = 4$ and $\nu_2 = 15$ is found to be 3.64. Is this value of *F* significantly different from zero at 5% level of significance ?

16. A sample of the monthly earnings records of 15 employees of company A has a variance of Rs. 15.90, while a similar sample of 87 employees for company B has a variance of Rs. 17.50. Is it safe to assume that there is less variance in company A than in company B ?

17. The nicotine contents in milligrams of two samples of tobacco were found to be as follows :

Sample A :	20	16	26	27	23	22	
Sample B :	27	33	42	35	32	34	38

Can it be said that two samples come from normal population having the same mean ?

18. Two laboratories carry out independent estimates of a particular chemical in a medicine produced by a certain firm. A sample is taken from each batch, halved and the separate halves sent to the two laboratories. The following data are obtained :

Number of samples	10
Mean value of the difference of estimates	0.6
Sum of the squares of the differences (from their means)	20

Is the difference significant ?

19. A fertiliser mixing machine is set to give 12 kg of nitrate for every quintal bag of fertiliser. Ten 100 kg bags are examined. The percentage of nitrate are as follows :

11, 14, 13, 12, 13, 12, 13, 14, 11, 12.

Is there reason to believe that the machine is defective ?

20. The marks obtained by two groups of students in a Statistics test are given below :

	Group A	*Group B*
No. of students	12	11
Mean Marks	42	38
Standard Deviation of Marks	10	15

On the basis of this data, can it be concluded that there is a significant difference in the mean marks obtained by the two groups ?

21. Two types of batteries are tested for their length of life and the following data are obtained :

	Size of sample	*Mean life in hours*	*Variance*
Type A	9	600	121
Type B	8	640	144

Is there a significant difference in the two means ?
[$t = 7.17$]

22. A correlation coefficient of 0.2 is found in a sample of 28 pairs. Use Z-test to find out if this is significantly different from zero. *(MBA, Delhi Univ., 2008)*

23. Two different types of drugs 'A' and 'B' were tried on certain patients for increasing weight, 6 persons were given drug A and 8 persons were given drug B. The increase in pounds is given below :

Drug A :	7	10	13	12	4	8		
Drug B :	12	8	3	18	16	9	8	3

Do the two drugs differ significantly with regard to their effect in increase in weight ?
[$t = 0.42$]

24. The mean weekly sales of the chocolate bar in general stores was 146.3 bars per store. After an advertising campaign, the mean weekly sales in 22 stores for a typical week increased to 157.7 bars and showed a standard deviation of 17.2. Was the advertising campaign successful ? *(MBA, Hyderabad Univ., 2005)*

25. A company desires to compare the effects on cavities of its brand A with a competitor's brand B. To eliminate some of the variation in test population pairs of identical twins are used. A brand is randomly assigned to each twin and is used for two years. The number of cavities developed during the period are reported below :

Pair :	1	2	3	4	5
Brand A :	4	5	7	6	3
Brand B :	3	5	4	2	1

Test at 5% level of significance, whether the data indicate a difference in cavities developed between the two brands.

26. Calculate the value of t and test the hypothesis of the difference between the average proteins for the two States as given below :

	Protein results						
State I :	12.6	13.4	11.9	12.8	13.0		
State II :	13.1	13.4	12.8	13.5	13.3	12.7	12.4

27. As a part of an industrial training programme, some trainees are instructed by method A, which is straight teaching machine instruction, and some are instructed by method B, which involves the personal attention of the instructor. The trainees instructed by each method, and the scores they obtained in an appropriate achievement test are :

Method A :	71	75	65	69	73	66	68	71	74	68
Method B :	72	77	84	78	69	70	77	73	65	75

Test the claim that method B is more effective. Use 5% level of significance.

[$t = 1.977$]

28. Two independent samples of 8 and 7 items respectively gave the following values :

Sample A :	9	11	13	11	15	9	12	14
Sample B :	10	12	10	14	9	8	10	

Examine, whether the difference between the means of the two samples is significant ?

29. To test the effect of a fertiliser on rice production, 24 plots of land having equal areas were chosen. Half of these plots were treated with fertiliser and the other half were untreated. Other conditions were the same. The mean yield of rice on the untreated plots was 4.8 quintals with a standard deviation of 0.4 quintal, while the mean yield on the treated plots was 5.1 quintals with a standard deviation of 0.26 quintal. Can we conclude that there is significant improvement in rice production because of the fertiliser at 5% level of significance ?

[t=2.18]

30. To compare the efficiency of standard and electric typewriters, ten typists are chosen at random and trained in the use of both kinds of typewriters. They are then asked to type on each kind of typewriter for half an hour and their speeds measured average number of words typed per minute, are observed and given in the table below :

Typist :	A	B	C	D	E	F	G	H	I	J
Standard :	60	64	72	76	75	75	79	74	84	82
Electric :	55	62	70	90	70	72	78	70	90	100

Are you of the opinion that there is a vast difference in the efficiency of the two types of typewriters ?

Suppose your company decides to buy the standard typewriter for use only when the electric typewriter gives 20 words per minute greater than that of a standard typewriter. Based on the result obtained above, how should the company act ?

31. In an assignment, subjects were assigned at random between two conditions, five to each. Their scores are given below. Can one say that there is a significant difference between these two conditions ? What must be assumed in carrying out this test ?

Condition A :	128	115	120	110	103
Condition B :	123	115	130	135	113

32. A company selects 9 salesmen at random and their sales figures (in thousand Rs.) for the previous month are recorded. They then undergo a course devised by a business consultant and their sales figures for the following month are compared as shown in the table. Has the training course caused an improvement in the salesmen's ability ? Use 5% level.

Previous month :	75	90	94	85	100	90	69	70	64
Following month :	77	101	93	92	105	88	73	76	68

[$t = 3$]

33. Two random samples were drawn from two normal populations and their values are :

A :	66	67	75	76	82	84	88	90	92		
B :	64	66	74	78	82	85	87	92	93	95	97

Test, whether the two populations have the same variance at 5% level of significance.

[$F = 1.415$]

34. The lifetime of electric bulbs for a random sample of 10 from a large consignment gave the following data :

Sample :	1	2	3	4	5	6	7	8	9	10
Life in 1000 hours :	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that the average lifetime of bulbs is 4000 hours ?

35. A sample of 25 college students is given an aptitude test. The mean test score for the sample is 475 and the standard deviation is 30. It is believed that the students of the college are above the normal average test score of 470. Conduct an appropriate test at 5% level of significance.

36. A Drug is given to 10 patients, and the increments in their blood pressure were recorded to be 3, 6, –2, 4, –3, 4, 6, 0, 0, 2. Is it reasonable to believe that the drug has no effect on change of blood pressure ?

37. A machine is producing ball bearings with diameters of 0.5 inches. It is known that the standard deviation of the bearings is 0.005 inches. A sample of 25 ball bearings is selected and their average diameter is found to be 0.498 inches. Determine the 99 per cent confidence interval.

38. Ten students are selected at random from a college and their height (in cms.) are found to be 100, 104, 108, 110, 118, 120, 122, 124, 126 and 128 cms. In the light of these data, discuss the suggestion that the mean height of the students of the college is 110 cms. (Use 5% level of significance).

[$t = 1.936$]

39. For a random sample of 10 persons, fed on diet A, the increases in weight in pounds in a certain period were :

10 6 16 17 13 12 8 14 15 9

For another random sample of 12 persons, fed on diet B, the increases in weight in the same period were :

9 13 22 15 12 14 18 8 21 13 10 17

Test, whether the diets A and B differ significantly as regards their effect on increase in weight.

40. The wages of 10 workers taken at random from a factory are given below :

Wages (Rs.) : 1578, 1572, 1570, 1568, 1572, 1578, 1570, 1572, 1596, 1584.

Is it possible that the mean wage of all workers of this factory is Rs. 1580 ?

[t=1.481, yes]

41. Eight students were given a test in mathematics and after one month's coaching, they were given another test of the similar nature. The following table gives the increases in their marks in the second test over the first :

Roll No.	:	1	2	3	4	5	6	7	8
Increase in marks	:	4	–2	6	–8	12	5	–7	2

Do the marks indicate that the students have gained from coaching ?

42. 10 workers are selected at random from a large number of workers in a factory. The number or items produced by them on a certain day are found to be :

51, 52, 53, 54, 55, 56, 57, 58, 59, 60.

In the light of these data, would it be appropriate to suggest that the mean of the number of items produced in the population is 58 ?

[$t = 2.6$, yes]

43. An automatic device is set to fill 170 pills in each bottle of a certain medicine. A sample of ten bottles was taken. They were found to contain : 168, 164, 166, 167, 168, 169, 170, 170, 170 and 171 pills, with a standard deviation of 2.16 pills. Discuss whether the device is properly adjusted.

44. A random sample of 9 items is taken of a certain measurement. From the data, it is found that $\Sigma X = 108$ and $\Sigma X^2 = 1584$. Find the confidence limits for the population mean at 5% level of significance and test the hypothesis that the population mean is 8. (Table values of 't' at 5% level for 8 d.f. and 9 d.f. are respectively 2.306 and 2.262.)

45. The following data show weekly sales of a manufacturer before and after reorganisation of the sales function, 10 weeks from Sept. to Dec. in the successive years were selected for comparisons :

Week No.	1	2	3	4	5	6	7	8	9	10
Sales (before re-organisation) (in '000 Rs.)	15	17	12	18	16	13	15	17	19	18
Sales (after re-organisation) (in '000 Rs.)	20	19	18	22	20	19	21	23	24	24

Apply 't' test to determine whether reorganisation had any effect on sales.

46. Eleven sales executives trainees are assigned selling jobs right after their recruitment. After a fortnight, they are withdrawn from their field duties and given a month's training for executive sales. Sales executed by them in thousands of rupees before and after the training in the same period, are listed below :

Sales ('000 Rs.)												
(Before training)	:	23	20	19	21	18	20	18	17	23	16	19
Sales ('000 Rs.)												
(After training)	:	24	19	21	18	20	22	20	20	23	20	27

Do these data indicate that the training has contributed to their performance ?

(M.Com., DU, 1999; M. Com., MD Univ., 2001)

47. (*a*) Road testing of a random sample of cars was carried out to determine if mean mileage is greater for model I cars than for model II. The sample data for model I: $n_1 = 8$, $\bar{x}_1 = 26.0$ and $s_1 = 1.4$, model II: $n_2 = 10$, $\bar{x}_2 = 23.6$ and $s_2 = 1.2$. Specify the null hypothesis. Perform a hypothesis test at 5% level and interpret the results. *(MBA, M.D. Univ., 2006)*

(*b*) Two types of drugs were used on 5 and 7 patients for reducing their weights.

Drug A :	10	12	13	11	14		
Drug B :	8	9	12	14	15	10	9

Is there a significant difference in the efficacy of the two drugs ? If not which drug should you buy ?

(M. Com., Madurai Kamaraj Univ., 2007)

48. Two random samples gave the following results :

Sample	*Size*	*Sample mean*	*Sum of squares of deviations from mean*
1	10	15	90
2	12	14	108

Assuming normal population, test for the equality of population at 5% level of significance. *(MBA, IGNOU, 2002)*

49. Twelve children, each one selected from 12 sets of identical twins were trained by a certain method A and the remaining 12 children were trained by method B. At the end of the year, the following I. Q scores were obtained :

Pair :	1	2	3	4	5	6	7	8	9	10	11	12
MethodA :	124	118	127	120	135	130	140	128	140	126	130	126
MethodB :	131	127	135	128	137	131	132	125	141	118	132	129

Is this a sufficient evidence to indicate a difference in the average I. Q scores of the two groups ? *(MBA, Anna Univ., 2003)*

50. In a certain experiment to compare two types of food A and B, the following results of increase in weights are observed in subjects :

Subject→		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
Increase in weight	Food A	49	53	51	52	47	50	52	53	407
	Food B	52	55	52	53	50	54	54	53	423

Assuming that the two samples of subjects are independent, can we conclude that Food B is better than Food A in promoting weight gain ? *(MBA, IGNOU, 2006)*

51. A vending machine is supposed to discharge 8 ounces of coffee if the correct coins are inserted. To test whether the machine is operating properly, 16 cups of coffee are taken from the machine and measured. It is found that the mean and standard deviation of 16 measurements are 7.5 and 0.8 ounces respectively. Is the machine operating properly ?

(M. Com., Allahabad Univ., 2004)

52. Two designs A and B gave the following output in 9 trails of each, which is a better design. Why ?

	Output								
A :	16	16	53	15	31	17	14	30	20
B :	18	27	23	21	22	26	39	17	28

(MBA, Bharathidasan Univ., 2007)

53. Each day the major stock markets have a group of leading gainers in price (stocks that go up the most). On one day the standard deviation in the per cent change for a sample of 12 NASDAQ leading gainers was 16.8. On the same day, the standard deviation in the per cent change for a sample of 12 NYSE leading gainers was 8.9. Conduct a significance test for equal population variances to see whether it can be concluded that there is a difference in the volatility of the leading gainers on the two exchanges. What is your conclusion at 5% leve of significance?

(MBA, Delhi Univ., 2009)

*****

17

Chi-Square Test

INTRODUCTION

In the previous chapter on small sampling theory, it was necessary to make certain assumptions about the populations from which the samples were drawn. In many of the statistical tests, we had to assume that the samples came from normal populations. When this assumption cannot be justified, it is necessary to use procedures that do not require that these conditions be met. These procedures are generally referred to as non-parametric methods. In this chapter, we will discuss the χ^2 test which belongs to this category.

The χ^2 (pronounced as Chi-square) test is based on χ^2 distribution which was first used by Karl Pearson in the year 1900.

The Chi-square Distribution

For large sample size, the sampling (probability) distribution of χ^2 can be closely approximated by a continuous curve known as the Chi-square distribution. The probability function of χ^2 distribution is given by

$$F(\chi^2) = c(\chi^2)^{(v2)-1}\, e^{\chi^2/2}$$

where,

$e = 2.71828$

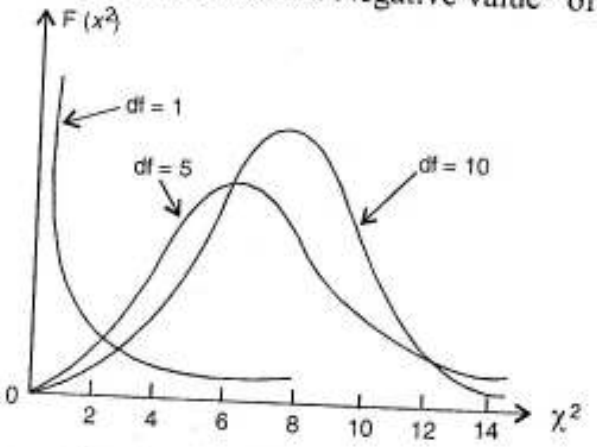
v = number of degrees of freedom

c = a constant depending only on v.

The Chi-square distribution has only one parameter v, the number of degrees of freedom. This is similar to the case of the t-distribution. Hence, $f(\chi^2)$ is a family of distributions, one for each value of v.

Important Properties of Chi-square Distribution

(1) χ^2 distribution is a continuous probability distribution which has the value zero at its lower limit and extends to infinity in the positive direction. Negative value* of χ^2 is not possible.



*The value of χ^2 can never be negative, since the differences between the observed and expected frequencies are always squared.

(2) The exact shape of the distribution depends upon the number of degrees of freedom v. For different values of v, we shall have different shapes of the distribution. In general, when v is small, the shape of the curve is skewed to the right and as v gets larger, the distribution becomes more and more symmetrical and can be approximated by the normal distribution.

(3) The mean of the χ^2 distribution is given by the degrees of freedom, *i.e.*, $E(\chi^2) = v$ and variance is twice the degrees of freedom, *i.e.*, $V(\chi^2) = 2v$.

(4) As v gets larger, χ^2 approaches the normal distribution with mean v and standard deviation $\sqrt{2v}$. In practice, it has been determined that the quantity $\sqrt{2\chi^2}$ provides a better approximation to normality than χ^2 itself for values of 30 or more. The distribution of $\sqrt{2\chi^2}$ has a mean equal to $\sqrt{2v-1}$ and a standard deviation equal to one.

(5) The sum of independent χ^2 variates is also a χ^2 variate. Therefore, if ${\chi_1}^2$ is a χ^2 variate with v_1 *d.f.* and ${\chi_2}^2$ is another χ^2 variate with v_2 *d.f.* independent of ${\chi_1}^2$, then their sum ${\chi_1}^2 + {\chi_2}^2$ is also a χ^2 variate with $v_1 + v_2$ *d.f.* This property is known as the additive property of χ^2.

Chi-square Test

The χ^2 test is one of the simplest and most widely used non-parametric tests in statistical work. It makes no assumptions about the population being sampled. The quantity χ^2 describes the magnitude of discrepancy between theory and observation, *i.e.*, with the help of χ^2 test we can know whether a given discrepancy between theory and observation can be attributed to chance or whether it results from the inadequacy of the theory to fit the observed facts. If χ^2 is zero, it means that the observed and expected frequencies completely coincide. The greater the value of χ^2, the greater would be the discrepancy between observed and expected frequencies. The formula for computing chi-square is :

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where, O = observed frequency

E = expected or theoretical frequency.

The calculated value of χ^2 is compared with the table value of χ^2 for given degrees of freedom at specified level of significance. If the calculated value of χ^2 is greater than the table value, the difference between theory and observation is considered to be significant, *i.e.*, it could not have arisen due to fluctuations of simple sampling. On the other hand, if the calculated value of χ^2 is less than the table value, the difference between theory and observation is not considered significant, *i.e.*, it could have arisen due to fluctuations of sampling.

The number of degrees of freedom is described as the number of observations that are free to vary after certain restrictions have been imposed on the data. For a uniform distribution, we place one restriction on the expected distribution—the total of sample observations.

In a contingency table, the degrees of freedom are calculated in a slightly different manner. The marginal total or frequencies place the limit on our choice of selecting cell frequencies. The cell frequencies of all columns but one $(c-1)$ and of all rows but one $(r-1)$ can be assigned arbitrarily and so the number of degrees of freedom for all cell frequencies is $(c-1)(r-1)$ where, c refers to columns and r refers to rows. Thus, in a 2×2 table, the degrees of freedom would be $(2-1)(2-1) = 1$ and in a 3×3 table, the degrees of freedom would be $(3-1)(3-1) = 4$.

Conditions for the Application of χ^2 Test

The following five basic conditions must be met in order for chi-square analysis to be applied :

(1) The experimental data (sample observation) must be independent of each other.

(2) The sample data must be drawn at random from the target population.

(3) The data should be expressed in original units for convenience of comparison and not in percentage or ratio form.

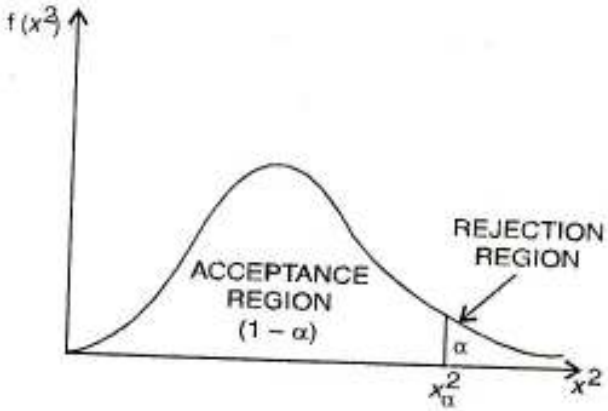
(4) The sample should contain at least 50 observations.

(5) There should not be less than five observations in any one cell (each data entry is known as a cell). For less than 5 observations, the value of χ^2 shall be over estimated and result in too many rejections of the null hypothesis.

Use of the Chi-square Table

To facilitate its many applications, the chi-square distribution has been extensively tabulated. The table of areas found in the Appendix gives value of χ^2 for various probabilities and various degrees of freedom. The value of α is given in the column headings, the degrees of freedom v are given in the rows and the body of the table gives the χ^2 values.

As depicted in the following figure, the value of χ^2 in the appendix table are given for various combinations of v and $1 - \alpha$.



Yates's Correction for Continuity

When using χ^2 analysis, it is important that a minimum of 80 per cent of the expected or theoretical frequencies in a cell be at least five and no cell have an expected frequency less than one. If the data results in expected frequencies less than five, wherever appropriate cell should be combined or the sample size should be increased until sufficient items fall into each cell.

The chi-square distribution is continuous distribution used with discrete data from a contingency table. When the expected frequencies are large, this approximate procedure is appropriate. In a 2×2 table, when expected frequencies are small, a correction was proposed by F. Yates in the year 1934 called "Yates's correction for continuity". The correction consists of :

$$\chi^2 \text{ (corrected)} = \frac{(|O_1 - E_1| - 0.5)^2}{E_1} + \frac{(|O_2 - E_2| - 0.5)^2}{E_2} + \ldots + \frac{(|O_k - E_k| - 0.5)^2}{E_k}$$

where $|O - E|$ means the absolute difference, ignoring plus and minus signs. Subtracting $\frac{1}{2}$ from the difference between O and E reduces the computed value of chi-square.

In general, the correction is made only when the number of degrees of freedom is $v = 1$. For large samples, this yields practically the same results as the uncorrected χ^2.

One problem with Yates's corrections should be noted. When the cells contain too few frequencies, ignoring Yates's correction might lead to excessive rejection of the hypothesis. On the other hand, Yates's correction tends to overcompensate for this and might result in excessive acceptance of the null hypothesis. The question is : what should be done by the analyst ? It may be reasonable to test the null hypothesis in the usual manner. If the hypothesis is accepted, we should be satisfied ; if rejection is indicated, then recalculate χ^2 using Yates's correction. Only if the null hypothesis is rejected without Yates's correction but accepted when the adjustment is used, should the analyst consider a more exact test than chi-square.

Grouping when Frequencies are Small

If small theoretical frequencies occur (less than 10 and certainly not less than 5), it is generally possible to overcome the difficulty by grouping two or more classes together. In other words, one or more classes with theoretical frequencies less than 5 may be combined into a single category before calculating the difference between observed and expected frequencies. The number of degrees of freedom would be determined with number of observations after the regrouping. This would be clear from the following :

Observed frequencies	: 364	376	218	89	33	13	2	1
Expected frequencies (based on Poisson dist.)	: 339	397	234	92	27	6	1	0

The last three classes should be combined together. After grouping, the position would be as follows :

Observed frequencies	: 364	376	218	89	33	16
Expected frequencies	: 339	397	234	92	27	7

The degrees of freedom would now be $8 - 2 - 1 = 5$. (Note the degrees of freedom are two less than the number of observations.)

Some important applications of χ^2-test are discussed in detail below :

(1) Sampling Distribution of the Sample Variance. The sampling distribution of the sample variance s^2 is particularly important in problems where one is concerned about the variability in a random sample. Since s^2 must always be positive, the distribution of s^2 cannot be a normal distribution.

The distribution of s^2 is a unimodal distribution which is skewed to the right, is a chi-square distribution. When the parent population is normal, with variance σ^2 and if random samples of size n with sample variance s^2 is drawn, can be shown to be related as :

$$s^2 = \frac{\chi^2\sigma^2}{v} = \frac{\chi^2\sigma^2}{n-1}$$

Therefore,
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

i.e., follows χ^2 distribution with $v = n - 1$.

(2) Confidence Interval for Variance. Confidence interval for variance σ^2 is based on the sampling distribution of $(n - 1)\ s^2/\sigma^2$ which follows χ^2 distribution with $v = (n - 1)$. A $100(1 - \alpha)$ per cent confidence interval for σ^2 is constructed by first obtaining an interval about $(n - 1)s^2/\sigma^2$. Two values of χ^2 are selected from the table (given in appendix) such that $\alpha/2$ is to the left of the smaller value and $\alpha/2$ is to the right of the larger values. Since the chi-square distribution is not symmetrical, $-\chi^2_{\alpha/2}$ does not give the approximate value of the left side of the distribution. The point that does give the correct probability is that of χ^2 cutting off $1 - \alpha/2$ of the right tail.

Therefore, a 100 $(1 - \alpha)$ per cent confidence interval for $(n - 1)\ s^2/\sigma^2$ is given by

$$-\chi^2_{\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < x^2_{1-\alpha/2}$$

Solving these inequalities for σ^2, we get

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

which is the required $100(1-\alpha)$ per cent confidence interval for σ^2.

(3) Tests of Hypothesis Concerning Variance. In testing hypothesis about the variance of a normally distributed population, the null hypothesis is $H_0 : \sigma^2 = \sigma_0^2$ where σ_0^2 is some specified value of the population variance.

We know that $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$

where s^2 is computed from a random sample of size n.

If $\chi^2 < \chi^2_{1-\alpha/2}$ and $\chi^2 > \chi^2_{\alpha/2}$, *i.e.*, when the computed value of χ^2 lies in the rejection region, we reject the null hypothesis, otherwise we accept the null hypothesis. This is shown in the diagram given below :

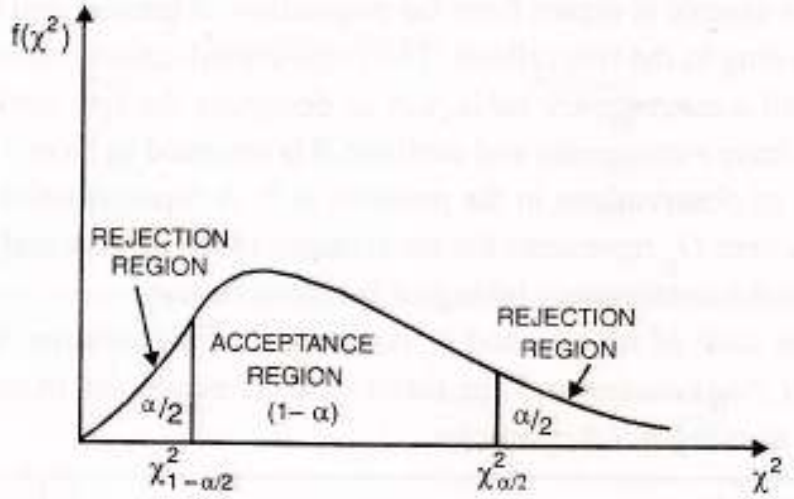


Illustration 1. Weights in kilograms of 10 shipments are given below :

38, 40, 45, 53, 47, 43, 55, 48, 52, 49.

Can we say that variance of the distribution of weight of all shipments from which the above sample of 10 shipments was drawn is equal to 20 square kilogram ?

Solution. Let the null hypothesis be that the variance of the distribution of shipments weight is 20 square kilogram, *i.e.*, $H_0 : \sigma^2 = 20$.

Weight (in kg.) X	$(X - \bar{X})$	$(X - \bar{X})^2$
38	−9	81
40	−7	49
45	−2	4
53	+6	36
47	0	0
43	−4	16
55	+8	64
48	+1	1
52	+5	25
49	+2	4
$\Sigma X = 470$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 280$

Sample mean $\bar{x} = \frac{\Sigma X}{n} = \frac{470}{10} = 47.$

Using the χ^2-test, statistic under the hypothesis $\sigma^2 = 20$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\Sigma(x-\bar{x})^2}{\sigma^2} = \frac{280}{20} = 14.$$

The table value of χ^2 for 9 *d.f.* at 5% level of significance is 16.919.

Since the calculated value of χ^2 is less than the tabulated value of χ^2, it is insignificant and the null hypothesis is accepted. Hence, we conclude that the data are consistent with the hypothesis that the variance of the distribution of weights of all shipments in the population is 20 kilograms.

(4) Test of Independence. One of the most frequent uses of χ^2 is for testing the null hypothesis that two criteria of classification are independent. They are independent if the distribution of one criterion in no way depends on the distribution of the other criterion. If they are not independent, there is an association between the two criteria. In the test of independence, the population and sample are classified according to some attributes. The test will indicate only, whether or not any dependency relationship exists between the attributes. It will not indicate the degree of association or the direction of the dependency.

To conduct the test, a sample is drawn from the population of interest and the observed frequencies are cross-classified according to the two criteria. The cross-classification can be conveniently displayed by means of a table called a *contingency table*. Let us designate the two attributes as A and B where, attribute A is assumed to have r categories and attribute B is assumed to have c categories. Furthermore, assume the total number of observations in the problem is N. A representation of these observations is shown below in a table where O_{ij} represents the observation in the ith row and jth column. Such a table in the matrix form is called a contingency table and is shown below.

In the table, R_i is the total of ith row and C_j is the total of jth column. The frequencies in these cells are termed as *cell frequencies* and the totals of the frequencies in each of the rows (R_i) and columns (C_j) are termed as *marginal frequencies*.

		Attribute B				Total
		B_1	B_2	B_3 ..... B_j	B_c	
Attribute A	A_1	O_{11}	O_{12}	O_{13} ..... O_{1j}	O_{1c}	R_1
	A_2	O_{21}	O_{22}	O_{23} ..... O_{2j}	O_{2c}	R_2
	A_3	O_{31}	O_{32}	O_{33} ..... O_{3j}	O_{3c}	R_3
	⋮	⋮	⋮	⋮ ⋮	⋮	⋮
	A_i	O_{i1}	O_{i2}	O_{i3} ..... O_{ij}	O_{ic}	R_i
	⋮	⋮	⋮	⋮ ⋮	⋮	⋮
	A_r	O_{r1}	O_{r2}	O_{r3} ..... O_{rj}	O_{rc}	R_r
	Total	C_1	C_2	C_3 ..... C_j	C_c	N

Expected cell frequencies are computed according to the multiplicative rule of probability. If two events are independent, the probability of their joint occurrence is equal to the product of their individual probabilities. Applying this rule to a contingency table, it is equivalent to say that, if two criteria of classification are independent, a joint probability is equal to the product of the two corresponding marginal probabilities. Thus, the expected cell frequencies are given by the formula :

$$E_{ij} = \frac{R_i}{N} \times \frac{C_j}{N} \times N = \frac{R_i C_j}{N}$$

To conduct the test, same χ^2 is employed as discussed earlier, *i.e.*,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad \text{or} \quad \sum \frac{(O - E)^2}{E}$$

will follow χ^2 distribution with $v = (r - 1)(c - 1)$ degrees of freedom.

While applying the test, the null hypothesis is that the two attributes are independent. If the calculated value of χ^2 is less than the table value at a specified level of significance, the null hypothesis holds true, *i.e.*, the two attributes are independent. If calculated value of χ^2 is greater than the table value, the null hypothesis is rejected, *i.e.*, the two attributes are associated.

Illustration 2. A sample of 200 persons with a particular disease was selected. Out of these, 100 were given a drug and the others were not given any drug. The results are as follows :

Number of Persons

	Drug	*No Drug*	*Total*
Cured	65	55	120
Not cured	35	45	80
Total	100	100	200

Test, whether the drug is effective or not.

Solution. Let us take the null hypothesis that the drug is not effective in curing the disease. Applying χ^2 test :

The expected* cell frequencies are computed as follows :

$$E_{11} = \frac{R_1 C_1}{N} = \frac{120 \times 100}{200} = 60; \quad E_{12} = \frac{R_1 C_2}{N} = \frac{120 \times 100}{200} = 60$$

$$E_{21} = \frac{R_2 C_1}{N} = \frac{80 \times 100}{200} = 40; \quad E_{22} = \frac{R_2 C_2}{N} = \frac{80 \times 100}{200} = 40$$

The table of expected frequencies is as follows :

60	60	120
40	40	80
100	100	200

O	E	$(O-E)^2$	$(O-E)^2/E$
65	60	25	0.417
35	40	25	0.625
55	60	25	0.417
45	40	25	0.625
			$\Sigma[(O-E)^2/E] = 2.084$

$$\chi^2 = \Sigma \frac{(O - E)^2}{E} = 2.084$$

*It may be noted that it is not necessary to calculate all the expected frequencies. It would be enough in a 2 × 2 table, if we calculate only one cell expected frequency. The others can be obtained by the process of deduction.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 2.084$$

$$v = (r-1)(c-1) = (2-1)(2-1) = 1$$

For $v = 1$, $\chi^2_{0.05} = 3.84$

The calculated value of χ^2 is less than the table value. The null hypothesis is accepted. Hence, the drug is not effective in curing the disease.

Illustration 3. A certain drug is claimed to be effective in curing cold. In an experiment on 500 persons with cold, half of them were given the drug and half of them were given the sugar pills. The patients' reactions to the treatment are recorded in the following table :

	Helped	*Harmed*	*No effect*	*Total*
Drug	150	30	70	250
Sugar pills	130	40	80	250
Total	280	70	150	500

On the basis of this data, can it be concluded that there is a significant difference in the effect of the drug and sugar pills ?

(MBA, Kumaun Univ., 2002; MBA, (HCA) DU, 2002)

Solution. Let us take the null hypothesis that there is no difference in the drug and sugar pills as far as their effect on curing cold is concerned.

Since it is a 2×3 table, the degrees of freedom would be $(2-1)(3-1) = 2$, *i.e.*, we will have to calculate only two expected frequencies and other four can be automatically determined.

Expected frequencies are computed as follows :

$$E_{11} = \frac{250}{500} \times 280 = 140; \quad E_{12} = \frac{250}{500} \times 70 = 35$$

The table of expected frequencies is :

140	35	75	250
140	35	75	250
280	70	150	500

Arranging the observed and expected frequencies in the following table :

O	E	$(O-E)^2$	$(O-E)^2/E$
150	140	100	0.714
130	140	100	0.714
30	35	25	0.714
40	35	25	0.714
70	75	25	0.333
80	75	25	0.333
			$\Sigma[(O-E)^2/E] = 3.522$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 3.522$$

The table of χ^2 for 2 *d.f.* at 5% level of significance is 5.99. The calculated value of χ^2 is less than the table value. Therefore the null hypothesis is accepted. Hence, we conclude that the drug and sugar pills do not differ significantly in curing cold.

(5) Test of Goodness of Fit. Tests of goodness of fit are used when we want to determine whether an actual sample distribution matches a known theoretical distribution. χ^2 test is popularly known as a test of goodness of fit for the reason, that it enables us to ascertain how well the theoretical distributions such as Binomial, Poisson, Normal, etc., fit empirical distribution, *i.e.*, those obtained from sample data. We hypothesize a theoretical distribution (Normal, for example) and then test to determine whether our sample came from or is comparable to the theoretical distribution. If there is a high degree of conformity

between the two distributions, any slight difference may be assumed to be the result of sampling variation. On the other hand, any large discrepancy between the two distributions may lead to the conclusion that the sample was drawn from some theoretical distribution other than the one proposed.

While applying the chi-square test of goodness of fit, the null hypothesis usually states that the sample is drawn from the theoretical population distribution, and the alternate hypothesis usually states that it is not. The following illustrations would illustrate the use of χ^2 test of goodness of fit.

Illustration 4. The number of parts for a particular spare part in a factory was found to vary from day to day. In a sample study, the following information was obtained :

Day	:	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Total
No. of parts demanded	:	1124	1125	1110	1120	1126	1115	6720

Test the hypothesis that the number of parts demanded does depend on the day of the week. *(MBA, Delhi Univ., 2000, 2005)*

Solution. Let us take the null hypothesis that the number of parts demanded does depend on the day of the week.

The number of spare parts demanded in a week are 6720 and if all days are same, we should expect 6720/6, *i.e.*, 1120 spare parts on a day of the week.

Day	O	E	$(O-E)^2$	$(O-E)^2/E$
Monday	1124	1120	16	0.014
Tuesday	1125	1120	25	0.022
Wednesday	1110	1120	100	0.089
Thursday	1120	1120	0	—
Friday	1126	1120	36	0.032
Saturday	1115	1120	25	0.022
				$\Sigma[(O-E)^2/E] = 0.179$

The table value of χ^2 for 5 *d.f.* at 5% level of significance is 11.07. The computed value of χ^2 is much less than the table value. The null hypothesis is accepted and we conclude that the demand for spare parts is dependent on the day of the week.

Illustration 5. 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? *(MBA, IGNOU., 2002)*

Solution. Let us take the null hypothesis on the assumption that male and female births are equally probable, the probability of a male birth is $p = 1/2$. The expected number of families can be calculated by the use of binomial distribution. The probability of x male births in a family of 5 is given by

$$f(x) = {}^5C_x p^x q^{5-x} \qquad [\text{for } x = 0,1,2,3,4,5]$$

$$= {}^5C_x (1/2)^5 \qquad [\because p = q = 1/2]$$

To get the expected frequencies, multiply $f(x)$ by the total number $N = 320$. The calculations are shown below in the table :

x	$f(x)$	*Expected frequency* $= N f(x)$
0	${}^5C_0\left(\frac{1}{2}\right)^5 = 1/32$	$320 \times 1/32 = 10$
1	${}^5C_1\left(\frac{1}{2}\right)^5 = 5/32$	$320 \times 5/32 = 50$
2	${}^5C_2\left(\frac{1}{2}\right)^5 = 10/32$	$320 \times 10/32 = 100$
3	${}^5C_3\left(\frac{1}{2}\right)^5 = 10/32$	$320 \times 10/32 = 100$
4	${}^5C_4\left(\frac{1}{2}\right)^5 = 5/32$	$320 \times 5/32 = 50$
5	${}^5C_5\left(\frac{1}{2}\right)^5 = 1/32$	$320 \times 1/32 = 10$

Arranging observed and expected frequencies in the following table and calculating χ^2 :

O	E	$(O-E)^2$	$(O-E)^2/E$
14	10	16	1.60
56	50	36	0.72
110	100	100	1.00
88	100	144	1.44
40	50	100	2.00
12	10	4	0.40
			$\Sigma[(O-E)^2/E] = 7.16$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 7.16$$

The table value of χ^2 for $v = 6 - 1 = 5$ at 5% level of significance is 11.07. The computed value of $\chi^2 = 7.16$ is less than the table value. Therefore, the null hypothesis is accepted. Thus, it can be concluded that male and female births are equally probable.

Illustration 6. The figures given below are (*a*) the theoretical frequencies of a distribution and (*b*) the frequencies of the distribution having the same mean, standard deviation and total frequency as in (*a*) :

(*a*)	1	12	66	220	495	792	924	792	495	220	66	12	1
(*b*)	2	15	66	210	484	799	943	799	484	210	66	15	2

Do you think that the normal distribution provides a good fit to the data ?

Solution. Let us take the null hypothesis that there is no difference in the observed frequencies and expected frequencies as obtained by the normal distribution.

Since the frequencies at the two corners are less than 5, they would be combined with the adjacent frequency.

O	E	$(O-E)^2$	$(O-E)^2/E$
1	2 }	16	0.941
12	15 }		
66	66	0	0.000
220	210	100	0.476
495	484	121	0.250
792	799	49	0.061
924	943	361	0.383
792	799	49	0.061
495	484	121	0.250
220	210	100	0.476
66	66	0	0.000
12	15 }		
1	2 }	16	0.941
			$\Sigma[(O-E)^2/E] = 3.839$

$v = 13 - 2 - 3 = 8$ (after grouping, 11 classes are left and for normal the degrees of freedom is less by 3 than the number of classes).

The table value of χ^2 for 8 *d.f.* at 5% level of significance is 15.51. The calculated value of χ^2 is less than the table value and hence the fit is good.

(6) Test of Homogeneity. It is frequently of interest to explore the proposition that several populations are homogeneous with respect to some characteristic of interest. For example, we may be interested in knowing of some raw material available from several retailers is homogeneous. Another way of stating the problem is to say that we are interested in testing the null hypothesis that several populations are homogeneous with respect to the proportion of subject falling into several categories or some other criterion of classification. A random sample is drawn from each

of the population and the number in each sample falling into each category is determined. The sample data is displayed in a contingency table. The analytical procedure is same as that discussed for test of independence.

The main difference is that, in tests of independence, we are concerned with the problem whether the two attributes are independent or not while in tests of homogeneity, we are concerned whether the different samples come from the same population. Another difference is that test of independence involve a single sample but test of homogeneity involves two or more samples, one from each population. When there are two populations involved, and when the characteristics of interest consist of two categories, the test of homogeneity is the same as testing hypothesis about the difference between two population's proportions which was discussed in the chapter on tests of hypothesis.

Illustration 7. A random sample of 400 persons was selected from each of three age groups and each person was asked to specify which of three types of TV programmes be preferred. The results are shown in the following table :

TYPES OF PROGRAMME

Age group	*A*	*B*	*C*	*Total*
Under 30	120	30	50	200
30–44	10	75	15	100
45 and above	10	30	60	100
Total	140	135	125	400

Test the hypothesis that the populations are homogeneous with respect to the types of television programme they prefer.

Solution. Let us take the null hypothesis that the populations are homogeneous with respect to different types of television programmes they prefer.

O	E	$(O-E)^2$	$(O-E)^2/E$
120	70.00	2500.00	35.7143
10	35.00	625.00	17.8571
10	35.00	625.00	17.8571
30	67.50	1406.25	20.8333
75	33.75	1701.56	50.4166
30	33.75	14.06	0.4166
50	62.50	156.25	2.5000
15	31.25	264.06	8.4499
60	31.25	826.56	26.4499
			$\Sigma[(O-E)^2/E] = 180.4948$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 180.495$$

The table value of χ^2 for 4 *d.f.* at 5% level of significance is 9.488.

The calculated value of χ^2 is much greater than the table value. We reject the null hypothesis and conclude that the populations are not homogeneous with respect to the type of TV programmes preferred.

Cautions while Applying χ^2 Test

χ^2 test is very popularly used in practice. However, it is unfortunate to find that the number of misuses of χ^2 test has become surprisingly large. The test must be used with greater care, keeping in mind the assumptions on which it is based. Some sources of error in the application of this test revealed by a survey of all papers published in the journal of *Experiment Psychology* are :

(*i*) Small theoretical frequencies.
(*ii*) Neglect of frequencies of non-occurrence.
(*iii*) Indeterminate theoretical frequencies.
(*iv*) Incorrect or questionable categorizing.
(*v*) Failure to equalize the sum of the observed frequencies and the sum of the theoretical frequencies.
(*vi*) Use of non-frequency data.

It should also be noted that χ^2 test is not the only non-parametric test. There are many other non-parametric tests that can be used in business decisions.

MISCELLANEOUS ILLUSTRATIONS

Illustration 8. Of the 1,000 workers in a factory exposed to an epidemic, 700 in all were attacked, 400 had been inoculated and of these, 200 were attacked. On the basis of this information, can it be said that inoculation and attack are independent ?

Solution : The given information can be put in a tabular form as follows :

	Inoculated	*Not inoculated*	*Total*
Attacked	200	500	700
Not attacked	200	100	300
Total	400	600	1000

Let us take the null hypothesis that inoculation and attack are independent. Applying χ^2 test, the expected frequency corresponding to first row and first column is $E_{11} = \frac{700 \times 400}{1000} = 280$. Table of expected frequencies would be as follows :

280	420	700
120	180	300
400	600	1000

O	E	$(O-E)^2$	$(O-E)^2/E$
200	280	6400	22.857
200	120	6400	55.333
500	420	6400	15.238
100	180	6400	35.556
			$\Sigma[(O-E)^2/E] = 128.984$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 128.984$$

For $\nu = 1, \chi^2_{0.05} = 3.84$

The calculated value of χ^2 is much greater than the table value. The null hypothesis is rejected. Hence, inoculation and attack are not independent.

Illustration 9. A sample analysis of examination results of 200 MBA's was made. It was found that 46 students have failed, 68 secured a third division, 62 secured a second division and rest were placed in the first division. Are these figures commensurate with the general examination result which is in the ratio of 2:3:3:2, for various categories respectively ? *(MBA, DU, 2002)*

Solution. Let us take the null hypothesis that there is no difference in the observed and expected results. On the basis of ratio 2 : 3 : 3 : 2, the expected number of students failing, getting third division, second division, and first division, should be

$\frac{200 \times 2}{10}$ = 40, 60, 60, 40 respectively.

Applying χ^2 test :

Category	O	E	$(O-E)^2$	$(O-E)^2/E$
Failed	46	40	36	0.900
Third Division	68	60	64	1.067
Second Division	62	60	4	0.067
First Division	24	40	256	6.400
				$\Sigma[(O-E)^2/E] = 8.434$

The table value of χ^2 for 3 *d.f.* at 5% level of significance is 7.81. The calculated value of χ^2 is greater than the table value. The null hypothesis does not hold true. Hence, the given results are not commensurate with the general examination results.

Illustration 10. An automobile manufacturing firm is bringing out a new model. In order to map out its advertising campaign, it wants to determine whether the model appeal depends on age group or not. The firm takes a random sample from persons attending a preview of the new model and obtained the results summarised below :

	AGE GROUPS				
Persons who	*Under 20*	*20–40*	*40–50*	*50 and over*	*Total*
Liked the car	146	78	48	28	300
Disliked the car	54	52	32	62	200
Total	200	130	80	90	500

Test, whether the model appeal and age groups are independent. *(MBA, DU, 2002)*

Solution. Let us take the null hypothesis that the model appeals equally to all the age groups, *i.e.*, model appeal does not depend on age groups.

$$E_{11} = \frac{300}{500} \times 200 = 120, \quad E_{12} = \frac{300}{500} \times 130 = 78, \quad E_{13} = \frac{300}{500} \times 80 = 48$$

The table of expected frequencies is :

120	78	48	54	300
80	52	32	36	200
200	130	80	90	500

O	E	$(O-E)^2$	$(O-E)^2/E$
146	120	676	5.633
54	80	676	8.450
78	78	0	–
52	52	0	–
48	48	0	–
32	32	0	–
28	54	676	12.519
62	36	676	18.778
			$\Sigma[(O-E)^2/E] = 45.38$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 45.38$$

The table value of χ^2 for 3 d.f. at 5% level of significance is 7.81. The calculated value of χ^2 is much greater than the table value. The null hypothesis is rejected. Hence, the model appeal depends on the age groups.

Illustration 11. The following figures show the distribution of digits in numbers chosen at random from a telephone directory :

Digit	:	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.

[MBA, IIT, Roorkee, 2000; M.Com., Madras Univ., 2009]

Solution. The null hypothesis is that the digits occur equally frequently in the directory.

The expected frequency for each of the digits, 0, 1, 2, ...9 is 10,000/10 = 1,000

Arranging the observed and expected frequencies in the following table :

O	E	$(O-E)^2$	$(O-E)^2/E$
1,026	1,000	676	0.676
1,107	1,000	11,449	11.449
997	1,000	9	0.009
966	1,000	1,156	1.156
1,075	1,000	5,625	5.625
933	1,000	4,489	4.489
1,107	1,000	11,449	11.449
972	1,000	784	0.784
964	1,000	1,296	1.296
853	1,000	21,609	21.609
			$\Sigma[(O-E)^2/E] = 58.542$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 58.542$$

The table value of χ^2 for $\nu = 10-1 = 9$ *d.f.* at 5% level of significance is 16.919. The computed value of χ^2 is much greater than the table value. The null hypothesis is rejected. Thus, it can be concluded that the digits are not uniformly distributed in the directory.

Illustration 12. The number of automobile accidents per week in a certain city were as follows :

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accident conditions were the same during this 10 week period ?

(MBA, Delhi Univ., 1999)

Solution. Let the null hypothesis be that the number of accidents per week in a certain city are consistent with the belief that the accident conditions were same during the ten-week period.

As the total number of accidents over the 10 week period are 100, according to the statement of the null hypothesis, these accidents should be uniformly distributed over the 10 week period. Therefore, the expected number of accidents per week is equal to 100/10 = 10.

Week	O	E	$(O-E)^2$	$(O-E)^2/E$
1	12	10	4	0.4
2	8	10	4	0.4
3	20	10	100	10.0
4	2	10	64	6.4
5	14	10	16	1.6
6	10	10	0	0.0
7	15	10	25	2.5
8	6	10	16	1.6
9	9	10	1	0.1
10	4	10	36	3.6
				$\Sigma[(O-E)^2/E] = 26.6$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 26.6$$

Table value of χ^2 at 5% level of significance for 9 *d.f.* is 16.819.

Since the calculated value of χ^2 is greater than the table value, therefore, the null hypothesis is rejected. Hence, we conclude that the accident conditions are not the same (uniform) over the 10 week period.

Illustration 13. In a certain sample of 2000 families, 1400 families are consumers of tea, out of 1800 Hindu families, 1236 families consume tea. Use chi-square test to test whether there is any significant difference between consumption of tea among Hindu and non-Hindu families. *(MBA, Madurai Kamaraj Univ., 2003)*

Solution. The above data can be conveniently arranged in the following table as :

	Hindus	*Non-Hindus*	*Total*
No. of families consuming tea	1236	164	1400
No. of families not consuming tea	564	36	600
Total	1800	200	2000

Let the null hypothesis be that the two attributes (consumption of tea and community) are independent.

The expected frequencies are computed as follows :

$$E_{11} = \frac{R_1 C_1}{N} = \frac{1400 \times 1800}{2000} = 1260; \quad E_{12} = \frac{R_1 C_2}{N} = \frac{1400 \times 200}{2000} = 140$$

The table of expected frequencies is :

1260	140	1400
540	60	600
1800	200	2000

Arranging the observed frequencies with the corresponding expected frequencies as given in the following table :

O	E	$(O-E)^2$	$(O-E)^2/E$
1236	1260	576	0.458
564	540	576	1.067
164	140	576	4.114
36	60	576	9.600
			$\Sigma [(O-E)^2/E] = 15.239$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 15.239$$

The table value of χ^2 for 1 *d.f.* at 5% level of significance is 3.841. Since the calculated value of χ^2 is much greater than the table value of χ^2, the null hypothesis is rejected. Hence, we conclude that the two communities differ significantly as regards to the consumption of tea.

Illustration 14. The following table gives the number of good and bad parts produced by each of three shifts in a factory :

Shift.	*Good*	*Bad*	*Total*
Day	900	130	1030
Evening	700	170	870
Night	400	200	600
Total	2000	500	2500

Is there any association between the shift and the quality of parts produced ?

(MBA, Kumaun Univ., 2000; MBA, Delhi Univ., 2005)

Solution. Let us take the null hypothesis that there is no association between the shift and quality of parts produced. The observed frequencies are :

Shift	*Good*	*Bad*	*Total*
Day	900	130	1030
Evening	700	170	870
Night	400	200	600
Total	2000	500	2500

The expected frequencies are computed as follows :

$$E_{11} = \frac{1030}{2500} \times 2000 = 824 \; ; \quad E_{21} = \frac{870}{2500} \times 2000 = 696$$

The table of expected frequencies is :

824	206	1030
696	174	870
480	120	600
2000	500	2500

O	E	$(O-E)^2$	$(O-E)^2/E$
900	824	5776	7.010
700	696	16	0.023
400	480	6400	13.333
130	206	5776	28.039
170	174	16	0.092
200	120	6400	53.333
			$\Sigma[(O-E)^2/E] = 101.83$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 101.83$$

For $v = 2, \chi^2_{0.05} = 5.991$.

Since the calculated value of χ^2 is much greater than the table value, the null hypothesis is rejected. On the basis of given data, we can, therefore, conclude that there is association between shift and the quality of parts produced.

Illustration 15. It has been stated that potential respondents are more likely to reply to questionnaire printed on light coloured paper than a dark coloured paper. Questionnaires were sent out on a random basis with the following results :

Colour used	*Response received*	*No response*	*Total*
Light	120	80	200
Dark	100	100	200
Total	220	180	400

Use an appropriate test at 5% level of significance to determine whether or not light colour paper yields better response.

Solution. Let us take the null hypothesis that the colour of the paper does not affect the response. Applying χ^2 test,

$$E_{11} = \frac{200}{400} \times 220 = 110; \quad E_{12} = \frac{200}{400} \times 180 = 90$$

The table of expected frequencies is :

110	90	200
110	90	200
220	180	400

O	E	$(O-E)^2$	$(O-E)^2/E$
120	110	100	0.909
100	110	100	0.909
80	90	100	1.111
100	90	100	1.111
			$\Sigma[(O-E)^2/E] = 4.04$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 4.04$$

For $v = 1$, $\chi^2_{0.05} = 3.84$.

The calculated value of χ^2 is greater than the table value. The null hypothesis is rejected. Hence, we may conclude that the colour of the paper may affect the response.

Illustration 16. A school bought a total of 500 colour television sets. Three different brands were purchased, and their repair records were kept for each set's operation. The data is given below :

Brand	*Number of Repairs*			
	0	1	2 or more	Total
A	143	70	37	250
B	90	67	43	200
C	17	13	20	50
Total	250	150	100	500

Is there a relationship between brand and number of repairs ?

Solution. Let us take the null hypothesis that there is no relationship between brand and number of repairs. Applying χ^2 test :

$$E_{11} = \frac{250}{500} \times 250 = 125 \; ; \quad E_{12} = \frac{250}{500} \times 150 = 75$$

$$E_{21} = \frac{200}{500} \times 250 = 100 \; ; \quad E_{22} = \frac{200}{500} \times 150 = 60.$$

The table of expected frequencies is :

125	75	50	250
100	60	40	200
25	15	10	50
250	150	100	500

O	E	$(O-E)^2$	$(O-E)^2/E$
143	125	324	2.592
90	100	100	1.000
17	25	64	2.560
70	75	25	0.333
67	60	49	0.817
13	15	4	0.267
37	50	169	3.380
43	40	9	0.225
20	10	100	10.000
			$\Sigma[(O-E)^2/E] = 21.174$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 21.174$$

For $\nu = 4$, $\chi^2_{0.05} = 9.488$.

The calculated value of χ^2 is more than the table value. The null hypothesis is rejected. Hence, there is a relationship between brand and number of repairs.

Illustration 17. The divisional manager of a retail chain believes the average number of customers entering each of the five stores in his division weekly is the same.

In a given week, a manager reports the following number of customers in their stores:

3000, 2960, 3100, 2780, 3160.

Test the divisional manager's belief at the 10 per cent level of significance. *(MBA, Delhi Univ., 2003)*

Solution. Let us take the null hypothesis that there is no significant difference in the number of customers entering each of the five stores. The number of customers entering each of the five stores is 15000, therefore, the expected frequency for each store is 15000/5 = 3000. Applying χ^2 test :

O	E	$(O-E)^2$	$(O-E)^2/E$
3000	3000	0	0
2960	3000	1600	0.533
3100	3000	10000	3.333
2780	3000	48400	16.133
3160	3000	25600	8.533
			$\sum [(O-E)^2/E] = 28.532$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 28.532$$

For $\nu = 4$, $\chi^2_{0.10} = 13.277$.

The calculated value of χ^2 is more than the table value. The null hypothesis is rejected. Hence, there is a significant difference in the number of customers entering each of the five stores.

Illustration 18. A die is thrown 150 times with the following results :

No. turned up	:	1	2	3	4	5	6
Frequency	:	19	23	28	17	32	31

Test the hypothesis that the die is unbiased.

Solution. Let us take the null hypothesis that there is no significant difference in the observed and expected frequencies in the throw of the die, *i.e.*, die is unbiased.

The expected frequencies for 1, 2, 3, etc. would be $\frac{150}{6} = 25$. Applying the χ^2 test :

O	E	$(O-E)^2$	$(O-E)^2/E$
19	25	36	1.44
23	25	4	0.16
28	25	9	0.36
17	25	64	2.56
32	25	49	1.96
31	25	36	1.44
			$\sum [(O-E)^2/E] = 7.92$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 7.92$$

For $\nu = 5$, $\chi^2_{0.05} = 11.07$.

The calculated value of χ^2 is smaller than the table value. The null hypothesis is accepted. Hence, the die seems to be unbiased.

Illustration 19. Use chi-square test to test if the two attributes in the following contingency table are independent :

	Training			
Performance	*Intensive*	*Good*	*Average*	*Total*
Above Average	100	150	40	290
Average	100	100	100	300
Poor	50	80	150	280
Total	250	330	290	870

(MBA, Punjab Univ., 2005)

Solution. Let us take the null hypothesis that the attributes performance and training are independent, *i.e.*, not associated. Applying χ^2 test :

$$E_{11} = \frac{290}{870} \times 250 = 83.33; \quad E_{12} = \frac{290}{870} \times 330 = 110;$$

$$E_{21} = \frac{300}{870} \times 250 = 86.21; \quad E_{22} = \frac{300}{870} \times 330 = 113.79.$$

The table of expected frequencies is :

83.33	110.00	96.67	290
86.21	113.79	100.0	300
80.46	106.21	93.33	280
250	330	290	870

O	E	$(O-E)^2$	$(O-E)^2/E$
100	83.33	277.89	3.335
100	86.21	190.16	2.21
50	80.46	927.81	11.53
150	110.00	1600.00	14.545
100	113.79	190.16	1.671
80	106.21	686.96	6.468
40	96.67	3211.49	33.221
100	100.00	0.00	0.000
150	93.33	3211.49	34.41
			$\Sigma [(O-E)^2/E] = 107.39$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 107.39$$

For $\nu = 4$, $\chi^2_{0.05} = 9.49$.

The calculated value of χ^2 is much greater than the table value. The null hypothesis is rejected. Hence, performance and training are associated.

Illustration 20. A cigarette company interested in the effect of sex on the type of cigarettes smoked and has collected the following data from a random sample of 150 persons :

Cigarette	Male	Female	Total
A	25	30	55
B	40	15	55
C	30	10	40
Total	95	55	150

Test, whether the type of cigarette smoked and sex are independent. *(MBA, Osmania Univ., 2006)*

Solution. Let us take the null hypothesis that there is no association between the type of cigarettes smoked and the sex. Applying χ^2 test :

$E_{11} = \frac{55}{150} \times 95 = 34.83, E_{21} = \frac{55}{150} \times 95 = 34.83, E_{12} = \frac{55}{150} \times 55 = 20.17$

The table of expected frequencies is :

34.83	20.17	55
34.83	20.17	55
25.34	14.66	40
95	55	150

O	E	$(O-E)^2$	$(O-E)^2/E$
25	34.83	96.63	2.774
40	34.83	26.73	0.767
30	25.34	21.72	0.857
30	20.17	96.63	4.791
15	20.17	26.73	1.325
10	14.66	21.72	1.482
			$\Sigma[(O-E)^2/E] = 11.996$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 11.996$$

For $\nu = 2, \chi^2_{0.05} = 5.99$

The calculated value of χ^2 is greater than the table value. The null hypothesis is rejected. Hence, type of cigarette smoked and sex are not independent.

Illustration 21. A certain drug was administered to 456 males out of a total of 720 in a certain locality to test its efficacy against typhoid. Relevant data is given below :

	Infection	*No Infection*	*Total*
Administered the drug	144	312	456
Not Administered	192	72	264
Total	336	384	720

(MBA, Sukhadia Univ., 2004)

Solution : Let us take the null hypothesis that there is no significant difference in the infection caused due to administration of drug or otherwise.

$$E_{11} = \frac{456}{720} \times 336 = 212.8$$

The table of expected frequencies is :

212.8	243.2	456
123.2	140.8	264
336	384	720

O	E	$(O-E)^2$	$(O-E)^2/E$
144	212.8	4733.44	22.24
192	123.2	4733.44	38.42
312	243.2	4733.44	19.46
72	140.8	4733.44	33.6
			$\Sigma[(0-E)^2/E] = 111.73$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 111.73$$

For $\nu = 1, \chi^2_{0.05} = 3.84$

The calculated value of χ^2 is much higher than the table value. Hence, the hypothesis is rejected. There is a significant difference in the infection caused due to administration of drug.

Illustration 22. Five coins are tossed 3,200 times and the number of heads appearing each time are noted. At the end, the following results were obtained :

No. of heads :	0	1	2	3	4	5
Frequency :	80	570	1100	900	500	50

Use chi-square test of goodness of fit to determine whether the coins are unbiased. *(MBA, Hyderabad Univ., 2006)*

Solution. Let the null hypothesis be that the coins are unbiased. If the coins are unbiased, then the distribution of heads will follow binomial distribution. Calculating the expected frequencies by using the formula $f(x) = {}^nC_x\, p^x\, q^{n-x}$.

The table of expected frequencies is :

No. of heads	$f(x) = {}^nC_x\, p^x\, q^{n-x}$	*Expected frequency* $= N f(x)$
0	$^5C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = 1\left(\frac{1}{2}\right)^5$	$3200 \times \frac{1}{32} = 100$
1	$^5C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4 = 5\left(\frac{1}{2}\right)^5$	$3200 \times 5 \times \frac{1}{32} = 500$
2	$^5C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = 10\left(\frac{1}{2}\right)^5$	$3200 \times 10 \times \frac{1}{32} = 1000$
3	$^5C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = 10\left(\frac{1}{2}\right)^5$	$3200 \times 10 \times \frac{1}{32} = 1000$
4	$\left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right) = 5\left(\frac{1}{2}\right)^5$	$3200 \times 5 \times \frac{1}{32} = 500$
5	$^5C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0 = 1\left(\frac{1}{2}\right)^5$	$3200 \times \frac{1}{32} = 100$

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
80	100	−20	400	4.0
570	500	+70	4,900	9.8
1,100	1,000	+100	10,000	10.0
900	1,000	−100	10,000	10.0
500	500	0	0	0.0
50	100	−50	2500	25.0
				$\Sigma(O-E)^2/E = 58.8$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 58.8.$$

For $v = 5$, $\chi^2_{0.05} = 11.07$.

Since the computed value of χ^2 is much greater than the table value of χ^2, therefore, we reject the null hypothesis. Hence, the coins are biased.

Illustration 23. A production supervisor is interested in knowing if the number of breakdowns on four machines is independent of the shift using the machines. Test this hypothesis based on the following sample information :

	Machine				
	A	*B*	*C*	*D*	*Total*
Morning	15	10	18	12	55
Evening	12	8	15	10	45
Total	27	18	33	22	100

Shift

(MBA, Delhi Univ., 2004, 2007)

Solution. Let us take the null hypothesis that the number of breakdowns is independent of the shift using the machines. The expected frequencies are :

$$E_{11} = \frac{55}{100} \times 27 = 14.85,\ E_{12} = \frac{55}{100} \times 18 = 9.90,\ E_{13} = \frac{55}{100} \times 33 = 18.15$$

The table of expected frequencies is :

14.85	9.90	18.15	12.10	55
12.15	8.10	14.85	9.90	45
27	18	33	22	100

O	E	$(O-E)^2$	$(O-E)^2/E$
15	14.85	0.0225	0.0015
12	12.15	0.0225	0.0019
10	9.90	0.0100	0.0010
8	8.10	0.0100	0.0012
18	18.15	0.0225	0.0012
15	14.85	0.0225	0.0015
12	12.10	0.0100	0.0008
10	9.90	0.0100	0.0010
			$\Sigma[(O-E)^2/E] = 0.0101$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 0.0101$$

For $\nu = 3$, $\chi^2_{0.05} = 7.81$

The calculated value of χ^2 is much less from the table value. The null hypothesis is accepted. Hence, the number of breakdowns is independent of the shift using the machines.

Illustration 24. Two sample polls of votes for two candidates A and B for a public office are taken, one from among residents of rural area and one from urban areas. The results are given below. Examine, whether the nature of the area is related to the voting preference in this election.

Votes for / *Area*	*A*	*B*	*Total*
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

(MBA., IGNOU, 2001; MBA, Anna Univ., 2003)

Solution. Let us take the null hypothesis that the nature of area is not related to the voting preference in this election. Applying χ^2 test :

$$E_{11} = \frac{1000}{2000} \times 1170 = 585$$

The table of expected frequencies is :

585	415	1000
585	415	1000
1170	830	2000

O	E	$(O-E)^2$	$(O-E)^2/E$
620	585	1225	2.094
550	585	1225	2.094
380	415	1225	2.952
450	415	1225	2.952
			$\Sigma[(O-E)^2/E] = 9.992$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 9.992$$

For $v = 1$, $\chi^2_{0.05} = 3.84$

The calculated value of χ^2 is greater than the table value. The null hypothesis is rejected. Hence, the nature of area is related to the voting preference.

Illustration 25. In setting sales targets, the marketing manager makes the assumption that order potentials are the same for each of the four sales territories. A sample of 200 sales data is given below :

Sales Territories			
I	*II*	*III*	*IV*
60	45	59	36

Should the manager's assumption be rejected [Given : the chi-square value at 5% level of significance for 3 degrees of freedom is 7.81] *(MBA, Delhi Univ., 2003)*

Solution. Let us take the null hypothesis that order potentials are the same for each of the four sales territories. Applying χ^2 test : A sample of 200 sales for four territories is given. Therefore, the expected sale for each territory is 200/4 = 50.

O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
60	50	10	100	2.00
45	50	–5	25	0.50
59	50	9	81	1.62
36	50	14	196	3.92
				$\Sigma[(O-E)^2/E] = 8.04$

The table value of χ^2 for 3 *d.f.* at 5% level of significance is 7.81 which is less than the calculated value of χ^2. The null hypothesis is rejected and we can conclude that marketing manager assumption is not justified. Hence, order potentials are not same for each of the four sales territories.

Illustration 26. A sample of parts provided the following table data on quality of parts by production shift :

Shift	*Number good*	*Number Defective*	*Total*
First	368	32	400
Second	285	15	300
Third	176	24	200
Total	829	71	900

Use five per cent level of significance to test the hypothesis that quality of parts is independent of the production shift. *(MBA, Delhi Univ. 2008)*

Solution. Let us take the null hypothesis that there is no significant difference between the quality of part produced and the production shift. Applying χ^2 test :

$$E_{11} = 368.44, E_{21} = 276.33$$

The table of expected frequencies is :

368.44	31.56	400
276.33	23.67	300
184.23	15.77	200
829	71	900

O	E	$(O-E)^2$	$(O-E)^2/E$
368	368.44	0.1936	0.0005
285	276.33	75.1869	0.2720
176	184.23	67.7329	0.3676
32	31.56	0.1936	0.0061
15	23.67	75.1689	3.1757
24	15.77	67.7329	4.2950
			$\Sigma[(0-E)^2/E] = 8.1169$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 8.1169$$

For $\nu = 2, \chi^2_{0.05} = 5.99$

The calculated value of χ^2 is greater than the table value. Hence, the null hypothesis is rejected. We, therefore, conclude that quality of part is not independent of the production shift.

Illustration 27. At a level of significance of 0.10, can we conclude that the following 400 observations follow a Poisson distribution ?

No. of arrivals (per hr.)	0	1	2	3	4	5 or more
No. of hours	20	57	98	85	78	62

(MBA, IGNOU, 2003)

Solution. To solve this question first, we have to calculate expected frequencies by applying Poisson distribution and then using χ^2 test of goodness of fit to conclude whether given observations follow the distributions or not.

Let us take the null hypothesis that the given data fits to Poisson distribution.

FITTING OF POISSON DISTRIBUTION

X	f	fX
0	20	0
1	57	57
2	98	196
3	85	255
4	78	312
5	62	310
	$N = 400$	$\Sigma fX = 1130$

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{1130}{400} = 2.825$$

Expected Frequencies as per Poisson law

$[e^{-m} = 0.06]$

$$NP(0) = e^{-m} \times N = 0.06 \times 400 = 24$$

$$NP(1) = NP(0) \times m = 24 \times 2.825 = 67.8$$

$$NP(2) = NP(1) \times \frac{m}{2} = 67.8 \times \frac{2.825}{2} = 95.77$$

$$NP(3) = NP(2) \times \frac{m}{3} = 95.77 \times \frac{2.825}{3} = 90.18$$

$$NP(4) = NP(3) \times \frac{m}{4} = 90.18 \times \frac{2.825}{4} = 63.69$$

$$NP(5) = NP(4) \times \frac{m}{5} = 63.69 \times \frac{2.825}{5} = 35.99$$

Applying χ^2 test by rounding off the expected frequencies :

O	E	$(O-E)^2$	$(O-E)^2/E$
20	24	16	0.667
57	68	121	1.779
98	96	4	0.042
85	90	25	0.278
78	64	196	3.062
62	36	676	18.778
			$\Sigma[(O-E)^2/E] = 24.606$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 24.606$$

For $\nu = 5, \chi^2_{0.10} = 9.24$

The calculated value of χ^2 is greater than the table value. Therefore, we reject the null hypothesis. Hence, Poisson distribution does not provide good fit to the given data.

Illustration 28. In setting sales targets, the marketing manager makes the assumption that order potentials are the same for each of the four sales territories. A sample of 200 sales data is given below :

Sales Territories			
I	*II*	*III*	*IV*
60	45	59	36

Should the manager's assumption be rejected. (*MBA, Delhi Univ., 2009*)

Solution. Let us take the null hypothesis that the order potentials are the same for each of the four sales territories. Hence, the expected sales target should be, *i.e.*, 50 in each sales territory. Applying χ^2 test :

O	E	$(O-E)^2$	$(O-E)^2/E$
60	50	100	2.00
45	50	25	0.50
59	50	81	1.62
36	50	196	3.92
			$\Sigma[(O-E)^2/E] = 8.04$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 8.04$$

For $\nu = 3, \chi^2_{0.05} = 7.81$

The calculated value of χ^2 is more than the table value. Hence, the null hypothesis is rejected. We, therefore, conclude that the order potential is not the same for each of the four sales territories.

Illustration 29. The following table gives the number of aircraft accidents that occurred during the various days of the week. Test, whether the accidents are uniformly distributed over the week.

Days	Mon.	Tue.	Wed.	Thurs.	Fri.	Sat.
No. of accidents	14	18	12	11	15	14

(*MBA, IGNOU, 2006*)

Solution. Let us take the null hypothesis that the accidents are uniformly distributed over the week. Applying χ^2 test :

Days	O	E	$(O-E)^2$	$(O-E)^2/E$
Mon.	14	14	0	0.000
Tue.	18	14	16	1.143
Wed.	12	14	4	0.286
Thurs.	11	14	9	0.643
Fri.	15	14	1	0.071
Sat.	14	14	0	0.000
	84			$\Sigma[(O-E)^2/E = 2.143$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 2.143$$

For $\nu = 5, \chi^2_{0.05} = 11.07$. The calculated value is much less than the table value. The null hypothesis is accepted. We, therefore, conclude that the accidents are uniformly distributed.

Illustration 30. One of the questions in a recent survey conducted by an Airline consultancy firm was "In the past 12 months, when travelling for business, what type of Airline ticket did you purchase most often?" The data obtained are shown in the following contingency table :

Type of ticket	*Type of Flight*		
	Domestic	*International*	*Total*
First class	29	22	51
Business/executive class	95	121	216
Economy class	518	135	653
	642	278	920

Using 5% level of significance test for the independence of type of flight and type of ticket. (*MBA, Delhi Univ., 2006*)

Solution. Let us take the hypothesis that the type of ticket and type of flight are independent. Applying χ^2 test, let us calculate the expected frequencies :

$$E_{11} = \frac{51}{920} \times 642 = 35.59 \approx 36$$

$$E_{21} = \frac{216}{920} \times 642 = 150.73 \approx 151$$

36	15	51
151	65	216
455	198	653
642	278	920

O	E	$(O-E)^2$	$(O-E)^2/E$
29	36	49	1.361
95	151	3136	20.768
518	455	3969	8.723
22	15	49	3.267
121	65	3136	48.246
135	198	3969	20.045
			$\Sigma[(0-E)^2/E] = 102.410$

$$v = (r-1)(c-1) = (3-1)(2-1) = 2$$

For $v = 2, \chi^2_{0.05} = 5.99$

The calculated value of χ^2 is much greater than the table value. The hypothesis is rejected. Hence the type of ticket and the type of flight are not independent.

Illustration 31. In a study of brand loyalty in the automative industry, new car customers were asked whethr the make of their new car was the same as the make of their previous car. The break down of 600 responses shows the brand loyalty for domestic, European and American cars.

Purchaser	*Domestic*	*European*	*American*	*Total*
Same make	125	55	68	248
Different make	140	105	107	352
	265	160	175	600

Test the hypothesis to determine whether brand loyalty is independent of the manufacturer. Use level of significance 5%. What is your conclusion? If a significant difference is found, which manufacturer appears to have the greatest brand loyalty ?

(MBA, Delhi Univ., 2009)

Solution. Let us take the hypothesis that the brand loyalty is independent of the manufacturer. Applying χ^2 test, calculation of expected values :

$$E_{11} = \frac{248}{600} \times 265 = 109.53 \approx 110$$

$$E_{21} = \frac{248}{600} \times 160 = 66.13 \approx 66$$

The table of expected frequencies shall be

110	66	72	248
155	94	103	352
265	160	175	600

O	E	$(O-E)^2$	$(O-E)^2/E$
125	110	225	2.045
140	155	225	1.452
55	66	121	1.833
105	94	121	1.287
68	72	256	3.556
107	103	16	0.155
			$\Sigma[(0-E)^2/E] = 10.328$

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = 10.328$$

$$v = (r-1)(c-1) = (3-1)(2-1) = 2$$

For $v = 2, \chi^2_{0.05} = 5.99$

The calculated value of χ^2 is much greater than the table value. Hence the hypothesis is rejected. Hence the brand loyalty is not independent of the manufacturer.

Illustration 32. The personnel department of IBM is doing a study about job satisfaction. A random sample of 375 employees was given a test designed to diagnose the level of job satisfaction. Each employee's salary was also recorded in the table below. Use an appropriate significance test to determine if salary and job satisfaction are independent at 5% level of significance.

Salary versus Job Satisfaction

Satisfaction	*Under $ 50000*	*$ 50000—$ 75000*	*Over $ 75000*	*Total*
High	30	30	20	80
Medium	100	85	30	215
Low	45	20	15	80
Total	175	135	65	375

(MBA, Delhi Univ., 2009)

Solution. The χ^2 test of significance would be appropriate in this case. Let us take the hypothesis that the salary and job satisfaction are independent. The expected frequencies are computed as follows :

$$E_{11} = \frac{80}{375} \times 175 = 37.33 \approx 37, \qquad E_{12} = \frac{80}{375} \times 135 = 28.8 \approx 29,$$

$$E_{21} = \frac{215}{375} \times 175 = 100.33 \approx 100, \qquad E_{22} = \frac{215}{375} \times 135 = 77.4 \approx 77.$$

The table of expected frequencies shall be :

37	29	14	80
100	77	38	215
38	29	13	80
175	135	65	375

O	E	$(O-E)^2$	$(O-E)^2/E$
30	37	49	1.324
100	100	0	0.000
45	38	49	1.289
30	29	1	0.034
85	77	64	0.831
20	29	81	2.793
20	14	36	2,571
30	38	64	1.684
15	13	4	0.308
			$\Sigma[(O-E)^2/E] = 10.834$

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = 10.834$$

For $v = 4, \chi^2_{0.05} = 9.488$

The calculated value of χ^2 is more than the table value. The hypothesis is rejected at 5% level and it is concluded that the salary and job satisfaction are not independent.

PROBLEMS

1-A : Answer the following questions, each question carries **one** mark:

(*i*) In contingency table, which of the following determines the degrees of freedom.

(*a*) $(r-1)(c-1)$, (*b*) $(r-1)(c+1)$, (*c*) $(r+1)(c+1)$, (*d*) $(r-1)(c+1)$ *(MBA, Madurai-Kamaraj Univ., 2003)*

(*ii*) Define χ^2 distribution.

(*iii*) What is χ^2 test of goodness of fit ? *(M. Com., M.K. Univ., 2002)*

(*iv*) Is χ^2 test a non-parametric test ?
(*v*) What are Yate's corrections ?
(*vi*) Explain how χ^2 test is used in the test of homogeneity.
(*vii*) How degrees of freedom are determined in testing of independence ?
(*viii*) What is additive property of χ^2 test ?
(*ix*) What precautions should be kept in mind while using the χ^2 test ?

1-B : Answer the following questions, each question carries **four** marks:
(*i*) Define χ^2 distribution. State the uses of χ^2 test.
(*ii*) Explain the characteristics of χ^2 test.
(*iii*) What is χ^2 distribution ? Describe the uses of χ^2 test ?
(*iv*) Explain χ^2 test of goodness of fit.
(*v*) Explain the significance of χ^2 distribution. *(MBA, Madras Univ., 2003)*

2. (*a*) What is χ^2 test ? What are its uses ? *(MBA, Sukhadia Univ., 2005)*
(*b*) What is χ^2 test ? Under what conditions is it applicable ? Point out its role in business decision-making.

3. What is chi-square test of independence ? What cautions are necessary while applying this test ?

4. Explain Yates's method of correction for small frequencies in contingency table.

5. What is χ^2 test of goodness of fit ? What cautions are necessary while applying this test ?

6. (*a*) What is χ^2 test ? Explain its important uses with the help of an example. *(MBA, UP Tech Univ., 2007)*
(*b*) What is χ^2 test ? Point out its applications. Under what conditions this test is applicable ?

7. Write short notes on :
(*i*) Yates's corrections for continuity, (*ii*) Degrees of freedom, (*iii*) Test of Goodness of fit.

8. Discuss the chi-square test of goodness of fit of theoretical distribution to an observed frequency distribution. State the conditions for the validity of chi-square test.

9. (*a*) Discuss the importance of χ^2 test. How is it used to test the association ?
(*b*) Describe the χ^2 test of significance and state the various uses to which it can be put.
(*c*) Show that the sum of two independent chi-square variables is also a chi-square variable. *(MBA, Anna Univ., 2003)*

10. 1000 families were selected at random in a city to test the belief that high income families usually send their children to public school and the low income families often send their children to government schools. The following results were obtained :

		School	
Income	*Public*	*Govt.*	*Total*
Low	370	430	800
High	130	70	200
Total	400	500	1000

Test, whether income and type of schooling are independent.
[$\chi^2 = 22.5$, No.]

11. A study is conducted of the volume of calls recieved on the switchboard of an insurance firm. A count is made of the number of incoming calls per minute for a sample of 120 minutes. The results of the study are shown below :

No. of calls per annum :	0	1	2	3	4	5
No. of minutes :	50	40	16	10	5	1

The statistician making the study believes that the incoming calls are distributed according to the Poisson distribution. Do you think his assumption is true ?

12. A survey was carried out in a state among the Doctors, belonging to the Kural Health Service cadre (500 doctors) and among the medical Education directorate cadre (300 teaching doctors). They were asked a question "would it be acceptable to you, if the govt. proposes to hire all the doctors on fixed period contractual basis?" The doctors were to answer either as "Acceptable" or "Not acceptable". There was no third answer category "undecided". The following was the data compiled in a cross tabulated format :

Doctors	*Acceptable*	*Not acceptable*	*Total*
Rural Cadre	195	305	500
Teaching Cadre	140	160	300
Total	335	465	800

Apply χ^2 test and test the null hypothesis. *(MBA, HCA, Delhi Univ., 2008)*

13. Two factories using material from the same supplier and closely controlled to an agreed specification, produce output for a given period classified into three quality grades as follows :

	Quality grades (Output in tonnes)			
Factory	*A*	*B*	*C*	*Total*
X	42	13	33	88
Y	20	8	25	53
Total	62	21	58	141

Do this output figures show a significant difference at the 5% level ?

184. The theory predicts the production of beans in the four groups *A, B, C* and *D* should be 9 : 3 : 3 : 1. In an experiment among 1,600 beans, the numbers in the four groups were 882, 313, 287, and 118. Does the experimental result support the theory ? [$\chi^2 = 4.72$]

185. The following data relate to the sales in a time of trade depression, a certain article in great demand. Do the data suggest that the sales are significantly affected by depression ?

District where sales are :	*Districts*	
	Not hit by depression	*Hit by depression*
Satisfactory	140	60
Not satisfactory	40	60

[$\chi^2 = 25$, yes]

(M.Com., GND Univ., 2006)

186. Based on information on 1,000 randomly selected fields about the tenancy status of the cultivators of these fields and use of fertilisers collected in an agroeconomic enquiry, the following classification was noted :

	Owned	*Rented*
Using fertilizer	416	184
Not using fertilizer	64	336

Would you conclude that owner-cultivators are more inclined towards the use of fertilizers?

[$\chi^2 = 273.5$, yes]

(M.Com., Osmania Univ., 2005)

187. 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*
X	86	60	44	10
Y	40	33	25	2

Would you say that the sampling techniques adopted by the two researchers are significantly different ?

[$\chi^2 = 1.1971$]

(MBA, Delhi Univ., 2006)

188. A random sample of size 20 from a normal population gives a sample mean of 42 and sample standard deviation of 6. Test the hypothesis that the population standard deviation is 9. Clearly state the alternative hypothesis you allow for and the level of significance adopted.

[$\chi^2 = 8.89$]

189. A manufacturer of TV sets was trying to find out what variables influenced the purchase of a TV set. Level of income was suggested as possible variable influencing the purchase of TV set. A sample of 500 households was selected and the information obtained is classified as shown below :

	Have TV Set	*Do not have TV Set*
Low income group	0	250
Middle income group	50	100
High income group	80	20

Is there evidence from the above data of a relation between ownership of TV sets and level of income ?

190. A book has 700 pages. The number of pages with various numbers of misprints is recorded below. At the 5% significance level, are the misprints distributed according to Poisson law ?

No. of misprints :	0	1	2	3	4	5	Total
No. of pages with misprints :	616	70	10	2	1	1	700

[$\chi^2 = 12.81$, No.]

(M.Com., Delhi Univ., 2004)

191. The following contingency table shows the classifications of 2,000 workers in a factory, according to the disciplinary action taken by the management and their promotional experience :

Disciplinary action	*Promotional Experience*	
	Promoted	*Not promoted*
Not-offenders	146	462
Offenders	54	1338

Test, whether the disciplinary action taken and promotional experience are independent.

[$\chi^2 = 1.227$]

22. Four machines *A, B, C* and *D* are used to manufacture certain machine parts which are classified as first grade, second grade and third grade. The quality control engineer wants to test whether the quality of the product from the four machines is same. Data collected is as follows :

	Machines				
Grade	*A*	*B*	*C*	*D*	*Total*
First	620	750	400	530	2,300
Second	130	200	140	130	600
Third	50	50	60	40	200
Total	800	1,000	600	700	3,100

[$\chi^2 = 31.89$]

23. A certain drug is claimed to be effective in curing colds. In an experiment on 328 people with cold, half of them were given the drug and half of them were given sugar pills. The patients' reactions to the treatment are recorded in the following table :

	Helped	*Harmed*	*No effect*
Drug	104	20	40
Sugar pills	88	24	52

Test the hypothesis that the drug is no better than sugar pills for curing colds.

24. From the following data, find out whether there is any relationship between sex and preference of colour :

Colour	*Male*	*Female*	*Total*
Green	40	60	100
White	35	25	60
Yellow	25	15	40
Total	100	100	200

[$\chi^2 = 8.17$, yes] *(M.Com., Punjab Univ., 2005)*

25. In a survey of 200 boys, of whom 75 were intelligent, 40 had skilled fathers; while 85 of the unintelligent boys had unskilled fathers. Do these figures support the hypothesis that skilled fathers have intelligent boys ?

[$\chi^2 = 8.89$] *(MBA, Delhi Univ., 2003)*

26. The figures given below are (*i*) the theoretical frequencies of a distribution and (*ii*) the frequencies of the normal distribution having the same mean, standard deviation and the total frequency as in (*i*)

(*i*)	1	5	20	28	42	22	15	5	2
(*ii*)	1	6	18	25	40	25	18	6	1

Apply χ^2 test of goodness of fit.

27. Four different drugs have been developed for a certain disease. These drugs are used under three different environment (it is assumed that the environment might affect efficacy of drugs). The number of cases of recovery from the disease per 100 people who have taken the drugs is tabulated as follows :

	Drugs			
Environment	A_1	A_2	A_3	A_4
I	19	8	23	8
II	10	9	12	6
III	11	10	13	16

Test, whether the drugs differ in their efficacy to treat the disease, also whether there is any effect of environment on the efficacy of disease.

28. 2,000 digits were selected at random from a set of tables. The frequencies of the digits were given as below :

Digit :	0	1	2	3	4	5	6	7	8	9
Frequency :	180	200	190	230	210	160	250	220	210	150

Use the chi-square test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which these were chosen.

29. The result of a certain survey shows that out of 50 ordinary shops of small size, 35 are managed by men of which 17 are in cities, 12 shops in villages are run by women. Can it be inferred that shops run by women are relatively more in villages than in cities ? Use chi-square test.

[$\chi^2 = 3.572$]

30. For 2×2 contingency table :

	A	not A
B	a	b
not B	c	d

Prove that the chi-square test for independence of the two attributes A and B gives :

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

31. In a group of 100 persons, 56 were tall and 44 short. Of those who were tall 30 acted as leaders, 16 as followers and the rest were unclassifiable. Among those who were short, 14 acted as leaders, 22 as followers and the rest were unclassifiable. Tabulate the data and find out whether or not there is significant association between height and leadership.

32. In a study of market penetration, the marketing division of a company selected random samples of 200, 150 and 300 consumers from three cities and obtained the data given below. Do the data indicate that the extent of market penetration in the three cities is independent of the consumers knowledge of the product ?

City	*Never heard of product*	*Group heard but did not buy*	*Bought it at least once*	*Total*
1	36	55	109	200
2	45	56	49	150
3	54	78	168	300
Total	135	189	326	650

33. The number of machine malfunctions per shift at a factory is recorded for 180 shifts and the following data are obtained :

No. of malfunctions :	0	1	2	3	4	5	6	Total
No. of shifts :	82	42	31	12	8	3	2	180

What is a reasonable probability model for this type of data ?

Test, if this model describe the data adequately.

34. Study the effectiveness of three teaching methods (A), (B) and (C) from the following table :

Aptitude	*Age*			
	Young	*Middle*	*Old*	*Total*
Low	82(A)	87(B)	80(C)	249
Middle	92(B)	82(C)	81(A)	255
High	90(C)	83(A)	88(B)	261
Total	264	252	249	765

Do the teaching methods significantly differ in effectiveness on aptitude ?

35. An automobile company gives you the following information about age groups and the liking for particular model of car which it plans to introduce :

	Age Group			
	Below 25	*25–50*	*Above 50*	*Total*
Persons who liked the car	45	30	25	100
Disliked the car	55	20	25	100
Total	100	50	50	200

On the basis of above data, can it be concluded that the model appeal is independent of the age group ? *(MBA, DU, 2004)*

$[\chi^2 = 3]$

36. Boys and girls were sampled from a school and tested for their mathematical skills. Their classification into well skilled and poorly skilled categories was as below :

	Mathematical Skills		
	Good	*Poor*	*Total*
Boys	50	10	60
Girls	20	20	40
Total	70	30	100

Apply χ^2 test to find whether boys are better in mathematical skills to girls.
[$\chi^2 = 12.7$, yes]

37. L. Chandra, salesman for D. Paper Company, has 5 accounts to visit per day. It is suggested that the variable sales by Mr. Chandra may be described by the binomial distribution, with the probability of selling each account being 0.3. Given the following observed distribution of Chandra's number of sales per day, can we conclude that the distribution does in face follow the suggested distribution ? Use the .05 significance level.

No. of sales per day :	0	1	2	3	4	5
Frequency of no. of sales :	20	65	42	14	6	3

(MFC, Delhi Univ., 2005)

38. You are given the distribution of the number of defective units produced in a single shift in a factory over 100 shifts. Would you say that the defective units follow a Poisson distribution?

No. of defective units :	0	1	2	3	4	5	6
No. of shifts :	4	14	23	23	18	9	9

39. Price of a basket of goods and services showed the following trend in up-country and mid-town markets :

	Increasing	*Not increasing*
Mid-town	56	31
Up-country	18	6

Show of the trends in up-country prices and in mid-town prices has any significant association.

40. "A sample of 300 students of Under-Graduate and 300 students of Post-Graduate classes of a University were asked to give their opinion towards the autonomous colleges. 190 of the Under-Graduate and 210 of the Post-Graduate students favoured the autonomous status."

Present the above fact in the form of a frequency table and test at 5% level, that opinions of Under-Graduate and Post-Graduate students on autonomous status of colleges are independent.

41. Calculate the expected frequencies for the following data presuming the two attributes, *viz.*, condition of home and condition of child as independent :

	Condition of Home	
Condition of Child	*Clean*	*Dirty*
Clean	70	50
Fairly clean	80	20
Dirty	35	45

Use chi-square test at 5% level to state whether the two attributes are independent.
[$\chi^2 = 24.64$] *(M.Com., Madurai-Kamaraj Univ., 2005)*

42. 1000 students at college level were graded according to their IQ and economic conditions of their home. Use χ^2 test to find out, whether there is any association between economic condition at home and I.Q.

	I.Q.		
Economic Condition	*High*	*Low*	*Total*
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

[$\chi^2 = 31.75$] *(MBA, Osmania, Univ., MBA, Kumaun Univ., 2008)*

43. The following table gives the number of car accidents that occurred during the various days of the week. Find, whether the accidents are uniformly distributed over the week.

Day :	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
No. of accidents :	14	16	8	12	11	9	14

[$\chi^2 = 4.165$] *(M.Com., M.D. Univ., 2005; MBA, Delhi Univ., 2005)*

44. What are the assumptions in carrying out test of independence of attributes through chi-square? Set up an appropriate hypothesis for the data given below and draw your conclusions through some suitable test of significance method.

	Level of Intelligence		
Family Status	*Dull*	*Average*	*Brilliant*
Lower Middle	20	35	25
Middle	40	70	30
Upper Middle	40	30	30

45. (*a*) A marketing agency gives following information about the age groups and their liking for a particular model which the company plans to introduce :

	Age group			
	Below 20	*20–39*	*40–59*	*Total*
Liked	125	420	60	605
Disliked	75	220	100	395
Total	200	640	160	1000

On the basis of the above data, can it be concluded that the model appeal is independent of the age group.
[$\chi^2 = 42.79$] (*MBA, Kumaun Univ., 2002*)

(*b*) A die was thrown 9,000 times and of these 3,220 yielded a 3 or 4. Is this consistent with the hypothesis that the die was unbiased ? (*MBA, Bharathidasan Univ., 2001*)

46. A random sample of 400 persons was selected from each of three age groups and each person was asked to specify which of three types of TV programmes be preferred. The results are shown in the following table :

	Table Programme			
Age Group	*A*	*B*	*C*	*Total*
Under 30	120	30	50	200
30–44	10	75	15	100
45 and above	10	30	60	100
Total	140	135	125	400

Test the hypothesis that the populations are homogeneous with respect to the types of television programme they prefer.
(*MBA,Guru Jambheshwar Univ., 2007*)

47. The following information is obtained concerning an investigation of 50 ordinary shops of small size :

	No. of Shops		
	in towns	*in villages*	*Total*
Run by Men	17	18	35
Run by Women	3	12	15
Total	20	30	50

Can it be inferred that shops run by women are relatively more in villages than in towns ? Use chi-square test.
[$\chi^2 = 0.121$] (*MBA, Madurai Kamaraj Univ., 2006*)

48. The number of analysis sum by three operators during different shifts is given below. Test the hypothesis that the performance of the operators is independent of shifts

	Operator		
	1	*2*	*3*
I	97	58	32
II	78	46	39

49. Fit a Poisson distribution to the following data and test for goodness of fit.

X :	0	1	2	3	4	5	6
f :	275	72	30	7	5	2	1

(*MBA, Anna Univ., 2007*)

*****

18

Analysis of Variance

INTRODUCTION

Earlier in Chapter 17 on chi-square test, we used the chi-square test to examine the difference between more than two proportions and to make inferences about whether such samples are drawn from populations each having the same proportion. In this chapter, we shall learn a technique known as *analysis of variance* to test for the significance of the difference between more than two sample means and to make inferences about whether our samples are drawn from the populations having the same mean. The "analysis of variance" procedure or "F-test" is used in such problems, where we want to test for the significance of the difference among more than two sample means. In fact, the technique of analysis of variance is one of the most powerful statistical methods developed by R.A. Fisher.

The analysis of variance originated in agrarian research and its language is thus loaded with such agricultural terms as *blocks* (referring to land) and *treatments* (referring to populations or samples which are differentiated in terms of varieties of seeds, fertilisers or cultivation methods). Today, analysis of variance finds application in every type of experimental design, in natural sciences as well as social sciences and has become a very broad and technical subject. The methods of analysis of variance are a fundamental part of planned research and the design of experiment, comparative studies are essential in judging the effects of new technology, procedures and policies. Though analysis of variance can be used in a number of ways, in this chapter, an attempt would be made to illustrate some business applications of this highly useful tool.

Assumptions in Analysis of Variance

The analysis of variance technique is based on the following assumptions :

(1) Each sample is drawn randomly from a normal population and the sample statistics tend to reflect the characteristics of the population.

(2) The populations from which the samples are drawn have identical means and variances, *i.e.*,

$$\mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_n$$
$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots\ldots = \sigma_n^2$$

In case we are not in a position to make these assumptions in a particular problem, the analysis of variance technique should not be used. In such cases, we should consider using a "Non-parametric (distribution-free) technique."

Computation of Analysis of Variance

The null hypothesis taken while applying analysis of variance technique is that the means of different samples do not differ significantly. The procedure followed in the analysis of variance would be explained separately for

(1) One-way classification, and

(2) Two-way classification.

However, irrespective of the type of classification, the analysis of variance is a technique of partitioning the total sum of squared deviations of all sample values from the grand mean and is divided into two parts—sum of squares between the samples and sum of squares within the samples. Individual observation in the same treatment samples, however, can differ from each other only because of chance variation, since each individual within the group receives exactly the same treatment.

ONE-WAY CLASSIFICATION

The term 'one-factor analysis of variance' refers to the fact that a single variable or factor of interest is controlled and its effect on the elementary units is observed. In other words, in one-way classification, the data are classified according to only one criterion. Suppose we have k independent random samples of $n_1, n_2, .., n_k$ observations from k populations. The population means are denoted by $\mu_1, \mu_2, ..., \mu_k$. The one-way analysis of variance is designed to test the null hypothesis :

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

i.e., the arithmetic means of the population from which the k samples are randomly drawn are equal to one another. The steps involved in carrying out the analysis are :

(1) Calculate the variance between the samples

The variance (sum of squares) between samples reflects the contribution of both different treatments and chance to inter-sample variability. Sum of squares is a measure of variation. The sum of squares between samples is denoted by SSB. For calculating variance between sample, we take the total of the square of the variations of the means of various samples from the grand mean and divide this total by the degrees of freedom. Thus, the steps in calculating variance between samples will be :

(*a*) Calculate the mean of each sample, *i.e.*, $\bar{X}_1, \bar{X}_2 ....., \bar{X}_k$

(*b*) Calculate the grand mean $\bar{\bar{X}}$. Its value is obtained as follows :

$$\bar{\bar{X}} = \frac{\bar{X}_1, + \bar{X}_2 + ... + \bar{X}_k}{n_1 + n_2 + ... + n_k}$$

(*c*) Take the difference between the means of the various samples and the grand mean ;

(*d*) Square these deviations and obtain the total which will give sum of squares between the samples; and

(*e*) Divide the total obtained in step (*d*) by the degrees of freedom. The degrees of freedom will be one less than the number of samples, *i.e.*, if there are 4 samples then the degrees of freedom will be $4 - 1 = 3$ or in general $v = k - 1$, where k = number of samples.

(2) Calculate the variance within the samples

The variance (sum of squares) within samples measures those inter-sample differences that arise due to chance only. It is denoted by SSW. For calculating the variance within the samples, we take the total of the sum of squares of the deviation of various items from the mean values of the respective samples and divide this total by the degrees of freedom. Thus, the steps in calculating variance within the samples will be :

(*a*) Calculate the mean value of each sample, *i.e.*, $\bar{X}_1, \bar{X}_2 ....., \bar{X}_k$,

(*b*) Take the deviations of the various observations in a sample from the mean values of the respective samples.

(*c*) Square these deviations and obtain the total which gives the sum of squares within the samples.

(*d*) Divide this total obtained in step (*c*) by the degrees of freedom, the degrees of freedom is obtained by deducting from the total number of observations, the number of samples, *i.e.*, $v = n - k$, where k refers to the number of samples and n refers to the total number of all the observations.

(3) Calculate the *F*-ratio

Calculate the F-ratio as follows :

$$F^* = \frac{\text{Variance between the samples}}{\text{Variance within the samples}} \text{ or } F = \frac{S_1^2}{S_2^2}.$$

F is always computed with the variance between the sample means as the numerator and the variance within the sample means as the denominator. The denominator is computed by combining the variance within the k samples into single measures.

(4) Compare the calculated value of *F*

Compare the calculated value of F with the table value of F for the given degrees of freedom at a certain critical level (generally we take 5% level of significance). If the calculated value of F is greater than the table value of F, it indicates that the difference in sample means is significant, *i.e.*, it could not have arisen due to fluctuations of random sampling or, in other words, the *samples do not come from the same population*. On the other hand, if the calculated value of F is less than the table value, the difference is not significant and hence could have arisen due to fluctuations of random sampling.

Illustration 1. As head of a department of a consumer's research organisation, you have the responsibility for testing and comparing lifetimes of four brands of electric bulbs. Suppose you test the lifetime of three electric bulbs of each of the four brands. The data is shown below, each entry representing the lifetime of an electric bulb, measured in hundreds of hours :

	Brand		
A	*B*	*C*	*D*
20	25	24	23
19	23	20	20
21	21	22	20

Can we infer that the mean lifetime of the four brands of electric bulbs are equal ? *(MBA, Univ. of Roorkee, 2000)*

Solution. The null hypothesis is that the mean lifetime of the four brands of electric bulbs are equal, *i.e.*,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Let $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ and $\bar{X}_4$ denote the mean lifetime of Brand *A*, *B*, *C* and *D* respectively and be the overall grand mean. Then,

X_1	X_2	X_3	X_4
20	25	24	23
19	23	20	20
21	21	22	20
$\bar{X}_1 = 20$	$\bar{X}_2 = 23$	$\bar{X}_3 = 22$	$\bar{X}_4 = 21$

and $$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{4} = \frac{20 + 23 + 22 + 21}{4} = \frac{86}{4} = 21.5.$$

The variance between samples can be computed as follows :

$\bar{X}$	$\bar{\bar{X}}$	$(\bar{X} - \bar{\bar{X}})$	$(\bar{X} - \bar{\bar{X}})^2$
20	21.5	−1.5	2.25
23	21.5	+1.5	2.25
22	21.5	+0.5	0.25
21	21.5	−0.5	0.25
			$\Sigma(\bar{X} - \bar{\bar{X}})^2 = 5.0$

* If this ratio is close to 1, there would be little cause to doubt the null hypothesis of equality to population means. On the other hand, if the variability between groups is large compared to the variability within groups, we would suspect, the null hypothesis to be false.

$$s_{\bar{x}}^2 = \frac{\Sigma\left(\bar{X} - \bar{\bar{X}}\right)^2}{k-1} = \frac{5.0}{4-1} = \frac{5}{3}$$

Put $$\sigma_{\bar{x}} = \frac{s_{\bar{x}}}{\sqrt{n}} \text{ or } s_{\bar{x}}^2 = n\,\sigma_{\bar{x}}^2 = 3 \times \frac{5}{3} = 5.$$

[Here, n represents the sample size and not the number of samples (k).]

Therefore, our first estimate of the population variance is based on the variance between the sample means and is given by

$$s_1^2 = 5.$$

The variance within samples can be computed as follows :

Brand A		*Brand B*		*Brand C*		*Brand D*	
X	$(X-\bar{X})^2$	X	$(X-\bar{X})^2$	X	$(X-\bar{X})^2$	X	$(X-\bar{X})^2$
20	0	25	4	24	4	23	4
19	1	23	0	20	4	20	1
21	1	21	4	22	0	20	1
$\bar{X} = 20$	$\Sigma(X-\bar{X})^2 = 2$	$\bar{X} = 23$	$\Sigma(X-\bar{X})^2 = 8$	$\bar{X} = 22$	$\Sigma(X-\bar{X})^2 = 8$	$\bar{X} = 21$	$\Sigma(X-\bar{X})^2 = 6$

Therefore,

Sample variance, $s_1^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{2}{2} = 1$; Sample variance, $s_2^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{8}{2} = 4$

Sample variance, $s_3^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{8}{2} = 4$; Sample variance, $s_4^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{6}{2} = 3.$

Therefore, the pooled estimate s^2 is given by

$$s^2 = \frac{s_1^2 + s_2^2 + s_3^2 + s_4^2}{4} = \frac{1+4+4+3}{4} = 3.$$

Thus, the second estimate of the population variance based on within the samples is given by

$$s_2^2 = 3$$

Therefore, $$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{s_1^2}{s_2^2} = \frac{5}{3} = 1.67.$$

From the F-table in the appendix, the table value of F for (3, 8) d.f. and at 5% level of significance is 4.07. Since the computed value of $F = 1.67$ is less than the table value of $F = 4.07$, therefore, we accept our null hypothesis. Hence, the difference is insignificant and we can infer that the average lifetime of different brands of bulbs are equal.

The Analysis of Variance Table

Since there are several steps involved in the computation of both the between and within sample variances, the entire set of results may be organised into an analysis of variance (ANOVA) table. This table is summarised and shown below :

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom d.f.*	*Mean Square MS*	*Variance Ratio F*
Between samples	SSB	$c-1$	$MSB = \frac{SSB}{c-1}$	$F = \frac{MSB}{MSW}$
Within samples	SSW	$n-c$	$MSW = \frac{SSB}{n-c}$	
Total	SST	$n-1$		

To use ANOVA table, it is convenient to use the following short-cut computational formulas :

$$\text{Between samples sum of squares} = SSB = \sum_{j=1}^{c} \frac{T_j^2}{n_j} = \frac{T^2}{N}$$

$$\text{Within samples sum of squares} = SSW = \sum_{j=1}^{c} \sum_{j=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{c} \frac{T_j^2}{n_j}$$

$$\text{Total sum of squares} = SST = \sum_{j=1}^{c} \sum_{j=1}^{n_j} X_{ij}^2 - \frac{T^2}{N}.$$

The format for the ANOVA table using the computational formulas is shown below :

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom d.f.*	*Mean Square MS*	*Variance Ratio F*
Between Samples	$SSB = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$	$c-1$	$MSB = \frac{SSB}{c-1}$	$F = \frac{MSB}{MSW}$
Within Samples	$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{k} \frac{T_j^2}{N_j}$	$n-c$	$MSW = \frac{SSW}{n-c}$	
Total	$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2 - \frac{T^2}{N}$	$n-1$		

To use ANOVA table, let us consider Illustration 1 again and see how it helps in computation.

In order to use the computational formulas, the following four quantities must be computed :

$$\sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2, T_j, \sum_{j=1}^{c} \frac{T_j^2}{n_j}, \text{ and } \frac{T^2}{N}.$$

To obtain these quantities, let us make the following table :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
20	400	25	625	24	576	23	529
19	361	23	529	20	400	20	400
21	441	21	441	22	484	20	400
$\Sigma X_1 = 60$	$\Sigma X_1^2 = 1202$	$\Sigma X_2 = 69$	$\Sigma X_2^2 = 1595$	$\Sigma X_3 = 66$	$\Sigma X_3^2 = 1460$	$\Sigma X_4 = 63$	$\Sigma X_4^2 = 1329$

T = sum of all the observations = $\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4$

$= 60 + 69 + 66 + 63 = 258$

or $$\frac{T^2}{N} = \frac{(258)^2}{12} = \frac{258 \times 258}{12} = 5547.$$

$$\sum_{j=1}^{4} \sum_{i=1}^{3} X_{ij}^2 = X_1^2 + X_2^2 + X_3^2 + X_4^2$$

$$= 1202 + 1595 + 1460 + 1329 = 5586$$

$$\sum_{j=1}^{4} \frac{T_j^2}{n_j} = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} = \frac{(60)^2}{3} + \frac{(69)^2}{3} + \frac{(66)^2}{3} + \frac{(63)^2}{3}$$

$$= \frac{1}{3}[3600 + 4761 + 4356 + 3969] = \frac{16686}{3} = 5562.$$

With this information, the ANOVA table for the electric bulb problem can be set as given below :

ANOVA TABLE : ONE-WAY CLASSIFICATION

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Square MS*	*Variance Ratio F*
Between Samples	$SSB = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$ $= 5562 - 5547 = 15$	$4 - 1 = 3$	$MSB = \frac{15}{3}$ $= 5$	$F = \frac{MSB}{MSW}$
Within Samples	$SSW = \sum_{j=1}^{c} \sum_{j=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{c} \frac{T_j^2}{n_j}$ $= 5586 - 5562 = 24$	$12 - 4 = 8$	$MSW = \frac{24}{8}$ $= 3$	$= \frac{5}{3}$ $= 1.67$
Total	$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2 - \frac{T^2}{N}$ $= 5586 - 5547 = 39$	$12 - 1 = 11$		

If we compare the F-ratio with the previous F-ratio obtained, we can see, of course, that the two methods (conceptual and computational) have produced exactly the same results. Therefore, it is recommended that the computational method be utilized for ANOVA table since these computations are generally less tedious and easy to perform.

Coding of data

If we add, subtract, multiply or divide the given data, the solution will not change. To show this, let us subtract all the observations from 20 in the electric bulb illustration. The coded data are given below.

X_1	X_2	X_3	X_4
0	+ 5	+ 4	+ 3
− 1	+ 3	0	0
+ 1	+ 1	+ 2	0

To compute different quantities, let us make the following table :

X_1	X^2_1	X_2	X^2_2	X_3	X^2_3	X_4	X^2_4
0	0	+ 5	25	+ 4	16	+ 3	9
− 1	1	+ 3	9	0	0	0	0
+ 1	1	+ 1	1	+ 2	4	0	0
$\Sigma X_1 = 0$	$\Sigma X^2_1 = 2$	$\Sigma X_2 = 9$	$\Sigma X^2_2 = 35$	$\Sigma X_3 = 6$	$\Sigma X^2_3 = 20$	$\Sigma X_4 = 3$	$\Sigma X^2_4 = 9$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 0 + 9 + 6 + 3 = 18$$

$$\frac{T^2}{N} = \frac{18 \times 18}{12} = 27$$

$$\Sigma X_{ij}^2 = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 = 2 + 35 + 20 + 9 = 66$$

$$\Sigma \frac{T_j^2}{n_j} = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} = \frac{(0)^2}{3} + \frac{(9)^2}{3} + \frac{(6)^2}{3} + \frac{(3)^2}{3}$$

$$= [0 + 81 + 36 + 9]/3 = 42.$$

ANOVA TABLE

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom d.f.*	*Mean Square MS*	*Variance Ratio F*
Between Samples	42 – 27 = 15	4 – 1 = 3	$MSB = \frac{15}{3} = 5$	$F = \frac{5}{3}$
Within Samples	66 – 42 = 24	12 – 4 = 8	$MSW = \frac{24}{8} = 3$	= 1.67
Total	66 – 27 = 39	12 – 1 = 11		

It may be noted that, we again get the value of *F*-ratio.

Readers are advised to use the following relation to further simplify calculations :

Total Variance = Variance between samples + Variance within samples.

Therefore, if we know any two values, the third can be automatically obtained. For example, variance within samples = Total Variance – Variance between samples.

Illustration 2. The Amit Merchandising Company wishes to test whether its three salesmen *A*, *B* and *C* tend to make sales of the same size or whether they differ in their selling ability as measured by the average size of their sales. During the last week, out of 14 sales, *A* made 5, *B* made 4 and *C* made 5 calls. The following are the weekly sales (in Rs. Thousand) record of three salesmen :

A	*B*	*C*
300	600	700
400	300	300
300	300	400
500	400	600
0	—	500

Test, whether the three salesmen's average sales differ in size. *(MBA, Bharathidasan Univ., 2007)*

Solution. Let us take the null hypothesis that there is no significant difference in the average sales volume of the three salesmen, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3$. In order to simplify calculations, let us divide each observation by 100 so that the coded data are :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
3	9	6	36	7	49
4	16	3	9	3	9
3	9	3	9	4	16
5	25	4	16	6	36
0	0			5	25
$\Sigma X_1 = 15$	$\Sigma X_1^2 = 59$	$\Sigma X_2 = 16$	$\Sigma X_2^2 = 70$	$\Sigma X_3 = 25$	$\Sigma X_3^2 = 135$

The sum of the sales of various samples

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 15 + 16 + 25 = 56$$

Correction factor $\frac{T^2}{N} = \frac{(56)^2}{14} = 224$

Total sum of squares $(SST) = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \frac{T^2}{N}$

$= (59 + 70 + 135) - 224 = 264 - 224 = 40$

Sum of squares between samples $(SSB) = \Sigma \frac{T_j^2}{n_j} - \frac{T^2}{N}$

$= \frac{(15)^2}{5} + \frac{(16)^2}{4} + \frac{(25)^2}{5} - 224 = 45 + 64 + 125 - 224 = 10$

Sum of squares within samples $(SSW) = SST - SSB = 40 - 10 = 30$

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	10	2	5
Within samples	30	11	2.73
Total	40	13	

$F_{(2, 11)} = \frac{5}{2.73} = 1.83$

The table value for $F_{(2, 11)}$ at 5% level of significance = 3.98. The calculated value of F is less than the table value*. Hence, we accept the null hypothesis and conclude that three salesmen do not differ significantly in their selling ability as measured by the average size of their sales.

Illustration 3. Four machines *A, B, C* and *D* are used to produce a certain kind of cotton fabrics. Samples of size 4 with each unit as 100 square metres are selected from the outputs of the machines at random, and the number of flaws in each 100 square metres are counted, with the following result.

A	*B*	*C*	*D*
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23

Do you think that there is a significant difference in the performance of the four machines ?

(*MBA, Kumaun Univ., 2006*)

Solution. Let us take the null hypothesis that the machines do not differ significantly in performance, i.e., $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
8	64	6	36	14	196	20	400
9	81	8	64	12	144	22	484
11	121	10	100	18	324	25	625
12	144	4	16	9	81	23	529
$\Sigma X_1 = 40$	$\Sigma X_1^2 = 410$	$\Sigma X_2 = 28$	$\Sigma X_2^2 = 216$	$\Sigma X_3 = 53$	$\Sigma X_3^2 = 745$	$\Sigma X_4 = 90$	$\Sigma X_4^2 = 2038$

$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 40 + 28 + 53 + 90 = 211$

Correction factor $= \frac{T^2}{N} = \frac{(211)^2}{N} = \frac{44521}{16} = 2782.56$

*It may be pointed out that computer software package include programs for performing analysis of variance calculations.

$$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \frac{T^2}{N}$$

$$= 410 + 216 + 745 + 2038 - 2782.56 = 3409 - 2782.56 = 626.44$$

$$SSB = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4} - \frac{T^2}{N}$$

$$= \frac{(40)^2}{4} + \frac{(28)^2}{4} + \frac{(53)^2}{4} + \frac{(90)^2}{4} - 2782.56$$

$$= 400 + 196 + 702.25 + 2025 - 2782.56 = 540.69$$

$$SSW = SST - SSB = 626.44 - 540.69 = 85.75$$

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	540.69	3	180.23
Within samples	85.75	12	7.15
Total	626.44	15	

$$F_{(3,12)} = \frac{180.23}{7.15} = 25.207.$$

The table value for $F_{(3,12)}$ at 1% level of significance is 5.95. The calculated value of F is greater than the table value. Hence, we reject the null hypothesis and conclude that there is a significant difference in the performance of the four machines.

Illustration 4. A random sample is selected from each of three makes of rope and their breaking strength (in pounds) are measured, with the following results :

I	*II*	*III*
70	100	60
72	110	65
75	108	57
80	112	84
83	113	87
	120	73
	107	

Test, whether the breaking strength of the ropes differ significantly.

Solution. Let us take the null hypothesis that the breaking strength of the ropes does not differ significantly, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3$. For simplifying calculations, let us take 80 as common. The coded data are given below :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
–10	100	20	400	–20	400
–8	64	30	900	–15	225
–5	25	28	784	–23	529
0	0	32	1024	+4	16
+3	9	33	1089	+7	49
		40	1600	–7	49
		27	729		
$\Sigma X_1 = -20$	$\Sigma X_1^2 = 198$	$\Sigma X_2 = 210$	$\Sigma X_2^2 = 6526$	$\Sigma X_3 = -54$	$\Sigma X_3^2 = 1268$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = -20 + 210 - 54 = 136.$$

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(136)^2}{18} = \frac{18496}{18} = 1027.56$$

$$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \frac{T^2}{N} = 198 + 6526 + 1268 - 1027.56 = 6964.44$$

$$SSB = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - \frac{T^2}{N} = \frac{(-20)^2}{5} + \frac{(210)^2}{7} + \frac{(-54)^2}{6} - 1027.56$$

$$= 80 + 6300 + 486 - 1027.56 = 5838.44$$

$SSW = SST - SSB = 6964.44 - 5838.44 = 1126$

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	5838.44	2	2919.22
Within samples	1126	15	75.07
Total	6964.44	17	

$$F_{(2,15)} = \frac{2919.22}{75.07} = 38.89$$

The table value for $F_{(2,15)}$ at 5% level of significance is 3.68. The calculated value of F is greater than the table value. The null hypothesis stands rejected. We, therefore, conclude that the breaking strength of the ropes differ significantly.

TWO-WAY CLASSIFICATION

In a one-factor analysis of variance explained earlier, the treatments constitute different levels of a single factor which is controlled in one experiment. There are, however, many situations in which the response variable of interest may be affected by more than one factor. For example, sales of a particular brand of cosmetics, in addition to being affected by the point of sale display, might also be affected by the price charged, the size and or location of the store or the number of competitive products sold by the store. Similarly, petrol mileage may be affected by the type of car driven, the way it is driven, road conditions and other factors in addition to the brand of petrol used.

When it is believed that two independent factors might have an effect on the response variable of interest, it is possible to design the test so that an analysis of variance can be used to test for the effects of the two factors simultaneously. Such a test is called two-factor (way) analysis of variance.

Thus, with the two-factor analysis of variance, we can test two sets of hypothesis with the same data at the same time.

We can plan an experiment in such a way as to study the effects of two factors in the same experiment. For each factor, there will be a number of classes or levels.

The procedure for analysis of variance is somewhat different from the one followed while dealing with problems of one-way classification. In a two-way classification, the analysis of variance table takes the following form :

ANOVA TABLE : TWO-WAY CLASSIFICATION

Source of Variation	*Sum of Squares*	*d.f.*	*Mean Square*
Between columns	SSC	$c - 1$	$MSC = SSC/(c-1)$
Between rows	SSR	$r - 1$	$MSR = SSR/(r-1)$
Residual	SSE	$(c-1)(r-1)$	$MSE = SSE/(c-1)(r-1)$
Total	SST	$rc - 1$	

SSC = Sum of squares between columns

SSR = Sum of squares between rows

SSE = Sum of squares for the residual

SST = Total sum of squares.

The sum of squares for the source "Residual"* is obtained by subtracting from the total sum of squares, the sum of squares between columns and rows.

The total number of degrees of freedom = $cr - 1$

where, c refers to columns and r refers to rows.

Number of degrees of freedom between columns = $(c - 1)$

Number of degrees of freedom between rows = $(r - 1)$

Number of degrees of freedom for residual = $(c - 1)(r - 1)$

The total sum of squares, sum of squares between columns and sum of squares between rows are obtained in the same way as before.

Residual = Total sum of squares – Sum of squares between columns – Sum of squares between rows.

Illustration 5. A company appoints four salesmen, *A, B, C* and *D*, and observes their sales in three seasons—summer, winter and monsoon. The figures (in lakhs) are given in the following table :

	Salesman				*Total*
Season	*A*	*B*	*C*	*D*	
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Total	90	93	81	96	360

Carry out an analysis of variance.

Solution. Let us take the null hypothesis that there is no significant difference between the sales of salesmen and that of seasons. The above data are classified according to criteria (*i*) salesman and (*ii*) season. In order to simplify calculations, we code the data by subtracting 30 from each figure. The data in the coded form are given below :

	Salesman				*Seasons Total*
Season	*A*	*B*	*C*	*D*	
Summer	+6	+6	–9	+5	+8
Winter	–2	–1	+1	+2	0
Monsoon	–4	–2	–1	–1	–8
	0	3	–9	6	Grand Total $T = 0$

Correction factor = $\frac{T^2}{N} = \frac{(0)^2}{12} = 0$ (number of observations or N is 12).

Sum of squares between columns (salesmen)

This will be obtained by squaring the salesmen totals, dividing each total by the number of observations included in it, adding these figures and then subtracting the correction factor from them. Thus, sum of squares between salesmen is

$$= \frac{(0)^2}{3} + \frac{(3)^2}{3} + \frac{(-9)^2}{3} + \frac{(6)^2}{3} - \frac{T^2}{N} = 0 + 3 + 27 + 12 - 0 = 42$$

Degrees of freedom = $v = (4 - 1) = 3$.

Sum of squares between rows (seasons)

This will be obtained by dividing the squares of the season totals by the number of observations that make up each total, adding all such figures and subtracting these from the correction factor. Thus, sum of squares between seasons is

$$= \frac{(8)^2}{4} + \frac{(0)^2}{4} + \frac{(-8)^2}{4} - \frac{T^2}{N} = 16 + 0 + 16 - 0 = 32$$

Degrees of freedom = $v = (3 - 1) = 2$.

* In these types of problems involving two-way classification, "Residual" is the measuring rod testing significance. It represents the magnitude of variations due to forces called 'chance'.

Total sum of squares

This will be obtained by adding the squares of all the observations in the table and subtracting the correction factor therefrom. Thus, total sum of squares is

$$= (6)^2 + (-2)^2 + (-4)^2 + (6)^2 + (-1)^2 + (-2)^2 + (-9)^2 + (1)^2 + (-1)^2 + (5)^2 + (2)^2 + (-1)^2 - \frac{T^2}{N}$$

$$= 210 - 0 = 210$$

$$v = (12 - 1) = 11$$

The above information is presented in the following table :

ANOVA TABLE

Source of Variation	*Sum of Squares*	*d.f.*	*Mean Square*
Between columns (salesmen)	42	3	14
Between rows (seasons)	32	2	16
Residual	136	6	22.67
Total	210	11	

To test the hypothesis that there is no significant difference between the sales of salesmen and of seasons or, in other words, the three independent estimates of variance are the estimates of variance of a common population.

Now, first compare the salesmen variance estimate with the residual variance estimate, thus,

$$F = \frac{\text{Greater Variance}}{\text{Smaller Variance}} = \frac{22.67}{14} = 1.62.$$

The table value of F for 3 and 6 degrees of freedom at 5% level of significance is 4.76. The calculated value is less than this and we conclude that the sales of salesmen do not differ significantly.

Now, let us compare the season variance estimate with the residual variance estimate, thus,

$$F = \frac{\text{Greater Variance}}{\text{Smaller Variance}} = \frac{22.67}{16} = 1.42.$$

The critical value of F for 2 and 6 degrees of freedom at 5% level of significance is 5.14. The calculated value is less than this and hence there is no significant difference in the seasons as far as the sales are concerned. Thus, the test shows that the salesmen and the seasons are alike so far as the sales are concerned.

Illustration 6. The following data represent the number of units of production per day turned out by 5 different workers using 4 different types of machines :

	Machine Type			
	A	*B*	*C*	*D*
Workers				
1	44	36	48	38
2	48	40	50	44
3	37	38	40	36
4	45	34	45	32
5	40	44	50	40

Test (*a*) Whether the mean productivity is the same for 4 different machine types.

(*b*) Whether the 5 workers differ with respect to mean productivity.

Solution. Let us take the null hypothesis that (*a*) the mean productivity is the same for four different machines and (*b*) the 5 workers do not differ with respect to mean productivity. To simplify calculations, let us deduct 40 from each value.

Workers	Machine Type				Total
	A	B	C	D	
1	+4	−4	+8	−2	+6
2	+8	0	+10	+4	+22
3	−3	−2	0	−4	−9
4	+5	−6	+5	−8	−4
5	0	+4	+10	0	+14
	+14	−8	+33	−10	$T = 29$

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(29)^2}{20} = 42.05$$

Sum of squares between machines (Column)

$$= \frac{(14)^2}{5} + \frac{(-8)^2}{5} + \frac{(33)^2}{5} + \frac{(-10)^2}{5} - \frac{T^2}{N}$$

$$= 39.2 + 12.8 + 217.8 + 20 - 42.05 = 289.8 - 42.05 = 247.75$$

$$\nu = (c-1) = (4-1) = 3$$

Sum of squares between workers (rows)

$$= \frac{(6)^2}{4} + \frac{(22)^2}{4} + \frac{(-9)^2}{4} + \frac{(-4)^2}{4} + \frac{(14)^2}{4} - \frac{T^2}{N}$$

$$= 9 + 121 + 20.25 + 4 + 49 - 42.05 = 161.20$$

$$\nu = (r-1) = (5-1) = 4$$

Total sum of squares

$$= (4)^2 + (8)^2 + (-3)^2 + (5)^2 + (-4)^2 + (-2)^2 + (-6)^2 + (4)^2 + (8)^2 + (10)^2 + (5)^2 + (10)^2 + (-2)^2 + (4)^2 + (-4)^2 + (-8)^2 - \frac{T^2}{N}$$

$$= 16 + 64 + 9 + 25 + 16 + 4 + 36 + 16 + 64 + 100 + 25 + 100 + 4 + 16 + 16 + 64 - 42.05 = 532.95$$

Residual = Total sum of squares − Sum of squares between machines − Sum of squares between workers

$$= 532.95 - 247.75 - 161.20 = 124$$

$$\nu = (c-1)(r-1) = 3 \times 4 = 12$$

ANOVA TABLE

Source of Variation	S.S.	d.f.	MS	Variation Ratio or F
Between Machines	247.75	3	82.583	$\frac{82.583}{13.78} = 5.99$
Between Workers	161.20	4	40.30	$\frac{40.30}{13.78} = 2.92$
Residual	124	12	10.33	
Total	532.95	19		

(*a*) For $\nu_{2,\,12}$, $F_{0.05} = 3.49$

Since the calculated value (5.99) is greater than the table value (3.49), the null hypothesis is rejected. Hence, the mean productivity is not the same for four different types of machines.

(*b*) For $\nu_{4,\,12}$, $F_{0.05} = 3.26$

The calculated value (2.92) is less than the table value (3.26). The null hypothesis holds true. Hence, the 5 workers do not differ with respect to mean productivity.

MISCELLANEOUS ILLUSTRATIONS

Illustration 7. We wish to test whether there are any differences in the performance of 3 brands of television sets. Samples of size $n = 5$ are selected from each brand and the frequency of repair during the first year of purchase is observed. The results are given below :

T.V. Brand

A_1	A_2	A_3
4	7	4
6	4	6
7	3	6
5	6	3
8	5	1

In view of the above data, can it be concluded that there is a significant difference between the three brands ?

Solution. Let us take the null hypothesis that there is no significant difference in the three brands of television sets with regard to their performance, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3$. Carrying out analysis of variance :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
4	16	7	49	4	16
6	36	4	16	6	36
7	49	3	9	6	36
5	25	6	36	3	9
8	64	5	25	1	1
$\Sigma X_1 = 30$	$\Sigma X_1^2 = 190$	$\Sigma X_2 = 25$	$\Sigma X_2^2 = 135$	$\Sigma X_3 = 20$	$\Sigma X_3^2 = 98$

$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 30 + 25 + 20 = 75$

$$\text{Correction Factor} = \frac{T^2}{N} = \frac{(75)^2}{15} = 375.$$

$$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \frac{T^2}{N} = 190 + 135 + 98 - 375 = 48$$

$$SSB = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - \frac{T^2}{N}$$

$$= \frac{(30)^2}{5} + \frac{(25)^2}{5} + \frac{(20)^2}{5} - 375 = 180 + 125 + 80 - 375 = 10$$

$$SSW = SST - SSB = 48 - 10 = 38.$$

ANOVA TABLE

Source of variation	*Sum of square*	*Degrees of freedom*	*Mean Square*
Between Samples	10	2	5.00
Within Samples	38	12	3.17
Total	48	14	

$$F_{(2, 12)} = \frac{5.0}{3.17} = 1.58.$$

The table value for $F_{(2, 12)}$ at 5% level of significance = 3.89. The calculated value of F is less than the table value. Hence, we accept the null hypothesis and conclude that there is no significant difference between the three brands of television sets.

Illustration 8. The following represent the number of units of production per day turned out by 4 different workers using 5 different types of machines :

MACHINE TYPES

Worker	*A*	*B*	*C*	*D*	*E*	*Total*
1	4	5	3	7	6	25
2	6	8	6	5	4	29
3	7	6	7	8	8	36
4	3	5	4	8	2	22
Total	20	24	20	28	20	112

On the basis of this information, can it be concluded that (*a*) the mean productivity is the same for different machines, (*b*) the workers don't differ with regard to productivity ?

Solution. Let us take the null hypothesis that (*a*) the mean productivity of different machines is same, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ and that (*b*) the 4 workers don't differ in respect of productivity. Carrying out analysis of variance :

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(112)^2}{20} = 627.2.$$

Sum of squares between machines :

$$= \frac{(20)^2}{4} + \frac{(24)^2}{4} + \frac{(20)^2}{4} + \frac{(28)^2}{4} + \frac{(20)^2}{4} - \text{C.F.}$$

$$= 100 + 144 + 100 + 196 + 100 - 627.2 = 640 - 627.2 = 12.8$$

$$v = (c - 1) = (5 - 1) = 4.$$

Sum of squares between workers :

$$= \frac{(25)^2}{5} + \frac{(29)^2}{5} + \frac{(36)^2}{5} + \frac{(22)^2}{5} - \text{C.F.}$$

$$= 125 + 168.2 + 259.2 + 96.8 - 627.2 = 649.2 - 627.2 = 22$$

$$v = (r - 1) = (4 - 1) = 3.$$

Total sum of squares :

$$= (4)^2 + (6)^2 + (7)^2 + (3)^2 + (5)^2 + (8)^2 + (6)^2 + (5)^2 + (3)^2 + (6)^2 + (7)^2 + (4)^2 + (7)^2 + (5)^2 + (8)^2 + (8)^2 + (6)^2 + (4)^2 + (8)^2 + (2)^2 - 627.2$$

$$= 692 - 627.2 = 64.8.$$

Residual = Total *SS* – *SS* between machines – *SS* between workers

$$= 64.8 - 12.8 - 22.0 = 30$$

$$v = (c - 1)(r - 1) = (5 - 1)(4 - 1) = 12$$

ANOVA TABLE

Source of variation	*SS*	*d.f.*	*Mean Square*	*Variance ratio 'F'*
Between machines	12.8	4	3.20	$\frac{3.2}{2.5} = 1.28$
Between workers	22.0	3	7.33	$\frac{7.33}{2.5} = 2.93$
Residual	30.0	12	2.50	
Total	64.8	19		

For $v_{(4, 12)}$, $F_{0.05} = 3.26$

The calculated value of *F* is less than the table value. Our null hypothesis is true. Hence, there is no significant difference in the mean productivity of five different machines.

For $v_{(3, 12)}$, $F_{0.05} = 3.49$

The calculated value of *F* is less than the table value. Our null hypothesis is true. Hence, there is no significant difference in the mean productivity of four different workers.

Illustration 9. Four salesmen were posted in different areas by a company. The number of units sold by them is given below :

A	20	23	28	29
B	25	32	30	21
C	23	28	35	18
D	15	21	19	25

On the basis of this information, can it be concluded that there is a significant difference in the performance of the salesmen.

Solution. Let us take the null hypothesis that there is no significant difference in the performance of the four salesmen.

Sample I *A*	*Sample II* *B*	*Sample III* *C*	*Sample IV* *D*
20	25	23	15
23	32	28	21
28	30	35	19
29	21	18	25
Total : 100	108	104	80
$\bar{X}$: 25	27	26	20

$$\bar{\bar{X}} = \frac{25 + 27 + 26 + 20}{4} = \frac{98}{4} = 24.5$$

VARIANCE BETWEEN SAMPLES

$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$	$(\bar{X}_4 - \bar{\bar{X}})^2$
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
1.00	25.00	9.00	81.00

Sum of squares between samples = 1 + 25 + 9 + 81 = 116.

VARIANCE WITHIN SAMPLES

$(X_1 - \bar{X}_1)^2$	$(X_2 - \bar{X}_2)^2$	$(X_3 - \bar{X}_3)^2$	$(X_4 - \bar{X}_4)^2$
25	4	9	25
4	25	4	1
9	9	81	1
16	36	64	25
54	74	158	52

Sum of squares within samples = 54 + 74 + 158 + 52 = 338

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	116	3	38.67
Within samples	338	12	28.17

$$F = \frac{38.67}{28.17} = 1.37$$

For $v_1 = 3$ and $v_2 = 12$, $F_{0.05} = 3.24$. The calculated value of F is less than the table value. The null hypothesis holds true. Hence, it cannot be concluded that there is a significant difference in the performance of the four salesmen.

Illustration 10. The following table gives the yields on 15 sample plots under three varieties of seed :

A :	20	21	23	16	20
B :	18	20	17	15	25
C :	25	28	22	18	32

Find out whether the average yield of land under different varieties of seed show significant differences.

Solution. Let us take the null hypothesis that the average yield of land under different varieties of seed do not differ significantly. Applying analysis of variance technique :

X_1	X_2	X_3
20	18	25
21	20	28
23	17	22
16	15	28
20	25	32
Total : 100	95	135
$\bar{X}$: 20	19	27

$$\bar{\bar{X}} = \frac{20 + 19 + 27}{3} = \frac{66}{3} = 22$$

VARIANCE BETWEEN SAMPLES

$(X_1 - \bar{\bar{X}})^2$	$(X_2 - \bar{\bar{X}})^2$	$(X_3 - \bar{\bar{X}})^2$
4	9	25
4	9	25
4	9	25
4	9	25
4	9	25
20	45	125

Sum of squares between samples = 20 + 45 + 125 = 190

VARIANCE WITHIN SAMPLE

$(X_1 - \bar{X}_1)^2$	$(X_2 - \bar{X}_2)^2$	$(X_3 - \bar{X}_3)^2$
0	1	4
1	1	1
9	4	25
16	16	1
0	36	25
26	58	56

Sum of squares within samples = 26 + 58 + 56 = 140

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	190	2	95
Within samples	140	12	11.67

$$F = \frac{95}{11.67} = 8.14$$

For v_1=2 and v_2=12, the table value of F is 3.88. Since the calculated value is more than the table value, the null hypothesis is rejected. Hence, the average yield of land under different varieties of seed differ significantly.

Illustration 11. Three varieties of coal were analysed by five chemists and the ash content in the varieties was found to be as under :

	Chemist				
Variety	*I*	*II*	*III*	*IV*	*V*
A	9	7	6	5	8
B	7	4	5	4	5
C	6	5	6	7	6

Do the varieties differ significantly in their ash content ?

Solution. Let us take the null hypothesis that there is no significant difference in the varieties with regard to the ash conten
Applying analysis of variance technique :

Sample I	*Sample II*	*Sample III*	*Sample IV*	*Sample V*
X_1	X_2	X_3	X_4	X_5
9	7	6	5	8
7	4	5	4	5
6	5	6	7	6
Total : 22	16	17	16	19
$\bar{X}$: 7.33	5.33	5.66	5.33	6.33

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5}{N} = \frac{7.33 + 5.33 + 5.66 + 5.33 + 6.33}{5} = \frac{29.98}{5} = 5.996 \text{ or } 6$$

VARIANCE BETWEEN SAMPLES

$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$	$(\bar{X}_4 - \bar{\bar{X}})^2$	$(\bar{X}_5 - \bar{\bar{X}})^2$
1.7689	0.4489	0.1156	0.4489	0.1089
1.7689	0.4489	0.1156	0.4489	0.1089
1.7689	0.4489	0.1156	0.4489	0.1089
$\Sigma(\bar{X}_1 - \bar{\bar{X}})^2$	$\Sigma(\bar{X}_2 - \bar{\bar{X}})^2$	$\Sigma(\bar{X}_3 - \bar{\bar{X}})^2$	$\Sigma(\bar{X}_4 - \bar{\bar{X}})^2$	$\Sigma(\bar{X}_5 - \bar{\bar{X}})^2$
= 5.3067	= 1.3467	= 0.3468	= 1.3467	= 0.3267

Sum of squares between samples

= 5.3067 + 1.3467 + 0.3468 + 1.3467 + 0.3267 = 8.6736.

VARIANCE WITHIN SAMPLES

X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$	X_3	$(X_3 - \bar{X}_3)^2$	X_4	$(X_4 - \bar{X}_4)^2$	X_5	$(X_5 - \bar{X}_5)^2$
9	2.789	7	2.789	6	0.116	5	0.109	8	2.789
7	0.109	4	1.769	5	0.436	4	1.769	5	1.769
6	1.769	5	0.109	6	0.116	7	2.789	6	0.109
	$\Sigma(X_1 - \bar{X}_1)^2$		$\Sigma(X_2 - \bar{X}_2)^2$		$\Sigma(X_3 - \bar{X}_3)^2$		$\Sigma(X_4 - \bar{X}_4)^2$		$\Sigma(X_5 - \bar{X}_5)^2$
	= 4.667		= 4.667		= 0.668		= 4.667		= 4.667

Sum of squares within samples

= 4.667 + 4.667 + 0.668 + 4.667 + 4.667 = 19.336.

TOTAL SUM OF SQUARES

X_1	$(X_1 - \bar{\bar{X}})^2$	X_2	$(X_2 - \bar{\bar{X}})^2$	X_3	$(X_3 - \bar{\bar{X}})^2$	X_4	$(X_4 - \bar{\bar{X}})^2$	X_5	$(X_5 - \bar{\bar{X}})^2$
9	9	7	1	6	0	5	1	8	4
7	1	4	4	5	1	4	4	5	1
6	0	5	1	6	0	7	1	6	0
	$\Sigma(X_1 - \bar{\bar{X}})^2$		$\Sigma(X_2 - \bar{\bar{X}})^2$		$\Sigma(X_3 - \bar{\bar{X}})^2$		$\Sigma(X_4 - \bar{\bar{X}})^2$		$\Sigma(X_5 - \bar{\bar{X}})^2$
	= 10		= 6		= 1		= 6		= 5

Total sum of squares = 10 + 6 + 1 + 6 + 5 = 28.

Total sum of squares = Sum of squares between samples + Sum of squares within samples

= 8.6736 + 19.336 = 28 01.

Hence, our calculations are correct.

ANOVA TABLE

Source of Variation	*SS*	*d.f.*	*Mean Square*
Between samples	8.6736	2	4.337
Within samples	19.336	12	1.6113
Total	28.01		

$$F = \frac{4.327}{1.6613} = 2.60$$

For $v_1 = 2$ and $v_2 = 12$, $F_{0.05} = 3.88$. The calculated value of F is less than the table value. Our hypothesis holds true. Hence, we conclude that there is no significant difference in the ash content of five different varieties.

Illustration 12. The three samples have been obtained from normal populations with equal variances. Test the hypothesis that the population means are equal.

SAMPLE

I	*II*	*III*
8	7	12
10	5	13
7	10	13
14	9	12
11	9	14

Solution. Let us take the null hypothesis that there is no significant difference in the means of three samples.

Sample	*I*	*II*	*III*
	8	7	12
	10	5	9
	7	10	13
	14	9	12
	11	9	14
Total	50	40	60
$\bar{X}$	10	8	12

$$\bar{\bar{X}} = \frac{10 + 8 + 12}{3} = \frac{30}{3} = 10.$$

VARIANCE BETWEEN SAMPLES

$(X_1 - \bar{\bar{X}})^2$	$(X_2 - \bar{\bar{X}})^2$	$(X_3 - \bar{\bar{X}})^2$
0	4	4
0	4	4
0	4	4
0	4	4
0	4	4
0	20	20

Sum of squares between samples = 0 + 20 + 20 = 40

VARIANCE WITHIN SAMPLES

$(X_1 - \bar{X})^2$	$(X_2 - \bar{X})^2$	$(X_3 - \bar{X})^2$
4	1	0
0	9	9
9	4	1
16	1	0
1	1	4
30	16	14

Sum of squares within samples = 30 + 16 + 14 = 60

ANOVA TABLE

Source of Variation	*SS*	*d.f.*	*MS*	*F*
Between samples	40	2	20	$\frac{20}{5} = 4$
Within samples	60	12	5	
Total	100	14		

For $\nu_1 = 2$ and $\nu_2 = 12$, the table value of F at 5% level of significance is 3.38. The calculated value of F is more than the table value. The hypothesis is rejected. Hence, the population means are not equal.

Caution while Applying Analysis of Variance Technique

The analysis of variance has been developed under a set of rigid assumptions as pointed out in the beginning of the chapter. Whenever, any of these assumptions is not met, the F-test cannot be employed to yield valid inferences. It is indeed fortunate that many economic and business experiments do conform to these assumptions. However, where departure from the premises exist, the analysis of variance may still be applied by way of *transformation.* Transformation refers to a process of transforming the original data into some other form, such as square roots, inverse sines of logarithms, before the analysis is made.

PROBLEMS

Answer the following questions, each question carries **one** mark:

(*i*) What is Analysis of Variance ?

(*ii*) The technique of analysis of variance was developed by .....

(*iii*) Define F-test. *(M. Com., M.K. Univ., 2001)*

(*iv*) Give two applications of analysis of variance.

(*v*) On what assumptions, analysis of variance is based ?

(*vi*) What do you understand by one-way analysis of variance ?

(*vii*) Give the format of ANOVA table in one-way classification.

(*viii*) What is two-way classification in analysis of variance ?

(*ix*) Give the components of source of variation in one-way classification.

(*x*) What is coding method in analysis of variance ?

Answer the following questions, each question carries **four** marks:

(*i*) Explain one-way classification technique in analysis of variance. *(M. Com., M.K. Univ., 2002)*

(*ii*) Tabulate the ANOVA table in one-way classification. *(M. Com., M.K. Univ., 2001)*

(*iii*) Explain the F-test. What are the assumptions of F-test ? *(M. Com., M.K. Univ., 2001)*

(*iv*) Differentiate between one-way and two-way classification by giving suitable example.

(*v*) Explain the procedure involved in ANOVA for testing of a hypothesis.

(*vi*) What is ANOVA? *(MBA, Madras Univ., 2002)*

What is 'analysis of variance' and where it is used ? Give two suitable examples.

How is analysis of variance technique helpful in solving business problems? Illustrate your answer with suitable examples. *(MBA, Kumaun Univ., 2000)*

Briefly describe the procedure followed in analysis of variance.

What are the basic and common assumptions made for analysis of variance ?

Distinguish between one-way and two-way classification models and explain the procedure followed for carrying out analysis of variance.

(*a*) Explain the meaning and significance of Analysis of Variance.

(*b*) State some applications of the analysis of variance.

(*c*) Explain the use of Analysis of variance (ANOVA) to check how good is the regression.

How is the F-distribution related to the Student's t-distribution and the chi-square distribution? What important hypothesis can be tested by the F-distribution?

9. In order to determine, whether there are significant differences in the durability of three makes of computer, samples of size $n = 5$ are selected from each make and the frequency of repair during the first year of purchase is observed. The results are as follow :

	Make	
A	*B*	*C*
5	8	7
6	10	3
8	11	5
9	12	4
7	4	1

In view of the above data, what conclusion can you draw ?

[$F = 5.34$, $F_{2, 12}$ at 5% level = 3.89]

10. A plastic manufacturer tests the tensile strength of different types of polythene material. A sample of three measurements is taken for each material type and data in pounds per square inch are as follows :

Type I	*Type II*	*Type III*
200	260	245
215	255	248
218	277	272

Determine, if the mean tensile strength of the three different types of materials differ significantly.

[$F = 16.30$, $F_{2, 6}$ at 5% level = 5.14, yes] *(MBA, Hyderabad Univ., 2005)*

11. The number of automobiles arriving at four toll stations were recorded for 2 hours time period (10 A.M. to 12 P.M.) for each of six different days. The data are as follows :

Day	*Gate I*	*Gate 2*	*Gate 3*	*Gate 4*
Monday	200	228	212	301
Tuesday	208	230	215	305
Wednesday	225	240	228	288
Thursday	223	242	224	212
Friday	228	210	235	215
Saturday	220	208	245	200

(*a*) Determine, whether the rate of arrival is essentially the same at each toll station.

(*b*) Determine, whether the rate of arrival differs significantly during the six different days of the week or not.

[(*a*) $F = 1.78$, No ; (*b*) $F = 0.56$, No]

12. Following table gives the number of refrigerators sold by 4 salesmen in three months:

		Salesmen		
Month	*A*	*B*	*C*	*D*
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

(*a*) Determine, whether there is any significant difference in the average sales made by four salesmen.

(*b*) Determine, whether the sales differ with respect to different months.

[(*a*) $F = 1.01$, No ; (*b*) F=3.29, No]

13. Miss Neena, a supervisor has 3 typists working under her supervision. She is concerned with the time they spend on the tea in addition to the normal lunch tea break. Her observations recorded in minutes for each typist are as follows :

	Average time (minutes)									
A	25	18	30	32	35	37	19			
B	24	22	26	28	30	32	28	26		
C	28	20	27	19	29	35	30	23	27	32

Can the differences in average time that the three typists spend on tea break be explained by chance variation ?

14. Five different brands of tyres used by a car rental agency in the process of deciding the brand of tyre to purchase as standard equipment for their fleet, find that each of five tyres of each brand last the following number of kilometres (in '000s) :

Tyre Brand				
A	*B*	*C*	*D*	*E*
36	46	35	45	41
37	39	42	36	39
42	35	37	39	37
48	37	43	35	35
47	48	38	32	38

Test the hypothesis that the five different brands of tyres have identical average life.

15. It is suspected that four machines, each in a canning operation fills cans to different levels on the average. Random samples of cans produced from each machine were taken and the fill in ounces was measured. The results are tabulated below :

	Machine		
A	*B*	*C*	*D*
10.20	10.22	10.17	10.15
10.18	10.27	10.22	10.37
10.36	10.26	10.34	10.28
10.21	10.25	10.27	10.40
10.25			10.30

Do the machines appear to be filling the cans at different average levels ?

16. During the last week, there were 14 sales calls. *A* made 5 calls, *B* made 4 calls and *C* made 5 calls. Following are the weekly sales (in 000's Rs.) record of the three salesmen :

	Salesmen		
	A	*B*	*C*
Calls	3	6	7
	4	3	3
	3	3	4
	5	4	6
	0	–	5

With the help of analysis of variance, test the selling ability of the three salesmen. *(MBE, Delhi Univ., 2002)*

17. Suppose that we are interested in establishing the yield producing ability of four types of soyabeans, *A, B,* C and *D*. We have three blocks of land *X, Y* and *Z* which may be different in fertility. Each block of land is divided into four plots, and the different types of soyabeans are assigned to the plots in each block by a random procedure. The following results are obtained :

		Type		
Block	*A*	*B*	*C*	*D*
X	5	9	11	10
Y	4	7	8	10
Z	3	5	8	9

Test whether *A, B, C* and *D* are significantly different.

18. The chairman of a large chain of supermarkets was prepared to order a large number of frozen food display cases for use in the markets. Before placing the order, he decided to test the products by storing half-litre containers of milk in the case made by different manufacturers and observing the spoilage time of each types of case. The display case came from three different manufacturers designated, *A, B* and *C*. Nine half-litre containers of milk were randomly selected and assigned, three to each case. The response variable observed was the spoilage time in day. The data for this test is provided below :

Treatment		
A	*B*	*C*
7	8	7
5	4	8
9	7	10

Test for a significant difference in the effect of the display cases at 5% level of significance.

19. The following table gives the monthly sales (in thousand rupees) of a certain firm in three different States by your different salesmen :

		Salesmen		
States	*W*	*X*	*Y*	*Z*
A	10	8	8	14
B	14	16	10	8
C	18	12	12	14

State, whether the difference between sales affected by the four salesmen and difference between sales affected in three States are significant.

20. Four brands of tyres were tested for durability and wear on specially designed machines which simulate road conditions. Four tyres of each brand were subjected to the same test and the number of kilometres until wear out was noted for each tyre. The data in thousands of kilometres is provided below :

	Tyre Brand		
A	*B*	*C*	*D*
24	26	28	12
18	16	17	18
23	19	26	30
13	30	19	20

Test for a significant difference tyre mileage at the 5% level of significance.

21. Four different drugs have been developed for the cure of a certain disease. These drugs are tried on patients in three different hospitals. The results given below show the number of cases of recovery from the disease per 100 people who have taken the drugs. The randomized blocks design has been employed to eliminate the effects of the hospital.

	A	*B*	*C*	*D*
H_1	32	18	20	21
H_2	15	23	26	13
H_3	26	10	17	17

Carry out an analysis of variance and interpret your results.

22. A manufacturer of footballs wants to introduce two additional styles of footballs to accompany the plastic version he already produces. The new footballs will be made of leather and rubber. All three styles were test marketed in five different stores. The manufacturer wants to concentrate on producing the type that promises the most sales. Is there a difference in sales of three types of footballs in the five different stores ?

Store	*Rubber*	*Plastic*	*Leather*
1	550	600	450
2	720	700	300
3	680	750	520
4	600	800	380
5	650	550	250

What should the manufacturer do about marketing his footballs?

23. A manufacturer has just introduced a new product that will be sold in sizes : small, medium and large. Five salesmen are randomly selected from the sales force and given each of the three products to sell. The sales figure for one month are used to find out whether there is a difference in sales volume for the different sizes. The amounts sold by the five salesmen are as follows :

	Size		
Salesman	*Small*	*Medium*	*Large*
1	850	900	880
2	720	880	760
3	880	970	930
4	900	890	670
5	750	960	880

Using a 0.05 level of significance, determine whether there is a significantly difference in the amount sold by size. What is your marketing decision ?

24. An economist wishes to assess the effects of Factor *A* (education) with five levels and Factor *B* (occupation) with four levels upon a person's annual earnings. The following data have been obtained for 20 randomly chosen person :
$SSW = 8,00,000$; $SSB = 9,00,000$ and Total $SS = 20,00,000$.
(*a*) Construct a one factor ANOVA table, using education as the only treatment. At 5 per cent significance level, can you conclude that the treatment means differ? (*b*) Construct a two factor ANOVA table, using education for occupation. What can you conclude about the respective null hypothesis for identical mean incomes for education levels and occupation ?

25. Mr. Ram wants to build a service station on one of three locations. He measures the traffic passing each location for six days. The following are the average amounts of traffic per hour passing each location for each of the six days :

Day	*Location A*	*Location B*	*Location C*
1	75	85	90
2	78	94	118
3	65	90	125
4	76	68	70
5	88	74	81
6	98	87	80

Is there any significant difference in the amount of traffic passing the three locations ? Where would you advise Mr. Ram to build his service station?

26. A company selling coffee appoints four salesmen *A,B,C* and *D*. Observe their sales in 3 seasons: summer, winter and monsoon.

The figures (in lakhs of rupees) are given below :

	Salesman			
Season	*A*	*B*	*C*	*D*
Summer	30	25	33	20
Winter	28	26	31	35
Monsoon	32	30	32	32

Carry out an analysis of variance and comment on your results.

27. The numbers of automobiles arriving at four gasoline stations were recorded for four-hour period from 8 A.M. to 12 Noon, from Monday through Saturday. Determine, whether the rate of arrival is essentially same at all stations.

Day	*Stations*			
	1	*2*	*3*	*4*
Monday	49	53	48	53
Tuesday	45	51	46	51
Wednesday	51	47	53	49
Thursday	48	53	42	51
Friday	50	50	50	53
Saturday	48	51	47	54

28. Three machines in a workshop are equally efficient. To measure the efficiency of four operators, the data on the number of units produced per shift by each operator on different machines on randomly selected shifts has been collected as follows :

Operator	*Machine*		
	A	*B*	*C*
I	22	20	19
II	24	19	17
III	27	23	21
IV	23	24	18

Test at 5% level of significance, whether machine operators are equally efficient.

29. A large retailer must make a choice between three sales locations within a shopping complex.

The following data are traffic counts for a 7-day period :

Location *X* :	643	542	569	552	607	514	576
Location *Y* :	249	404	378	337	426	298	345
Location *Z* :	458	513	485	482	539	491	368

Is there is significant difference in the average traffic count at the three locations ?

30. The following table gives monthly sales (in thousand rupees) of a certain firm in three States by its four salesmen :

State	*Salesmen*			
	I	*II*	*III*	*IV*
A	6	5	5	8
B	8	9	6	5
C	10	7	8	7

Test, whether there is any significant difference (*i*) between sales by the firm salesmen, and (*ii*) between sales by the four salesmen, and (*iii*) between sales in the three States.

31. The following are the defective pieces produced by four operators working, in turn, on four different machines :

Machine	*Operator*			
	B_1	B_2	B_3	B_4
A_1	34	28	33	29
A_2	31	24	35	23
A_3	27	20	43	72
A_4	28	28	29	26

Perform analysis of variance at 0.05 level of significance to ascertain whether variability in production is due to variability in operators' performance or variability in machines' performance.

32. To study the performance of three detergents and three different water temperatures, the following 'Whiteness' readings were obtained with specially designed equipment :

Water Temperature	*Detergent A*	*Detergent B*	*Detergent C*
Cold water	57	55	67
Warm water	49	52	68
Hot water	54	46	58

Perform a two-way analysis of variance, using 5% level of significance.

33. Agricultural engineers conducted an experiment to assess the effects of three different fertilisers on the yields of mango trees. They planted 15 plots of equal size and treated them alike except for the type of fertiliser applied, Fertiliser *A* was applied to 4 plots, fertiliser *B* to 5 plots and fertiliser *C* to 6 plots. The following table shows the yields, in quintals, per plot. Do these data provide sufficient evidence to indicate a difference in the treatment effect ? Use 5% level of significance.

Fertiliser						
A	7	5	6	4		
B	8	5	5	5	2	
C	5	6	4	1	2	3

34. The performances of a class of 300 students in the subjects of Statistics and Finance were graded into four classes *A*, *B*,*C* and *D*. The table below gives the cross tabulation of the number of students by grades in each of the two subjects :

	Statistics			
Finance	*A*	*B*	*C*	*D*
A	12	12	10	6
B	16	25	12	7
C	18	21	14	17
D	4	12	9	5

Test at significance of 5% and 1%, whether the performance can be inferred as independent.

35. The following table gives the number of units of production per day turned out by four different employees, using four different types of machines :

	Type of machines			
Employee	M_1	M_2	M_3	M_4
E_1	40	36	45	30
E_2	38	42	50	41
E_3	36	30	48	35
E_4	46	47	52	44

Using analysis of variance (*i*) test the hypothesis that the mean production is the same for the four machines and (*ii*) test the hypothesis that the four employees do not differ with respect to mean productivity.

36. In order to evaluate four comparable typewriters of different brands, five typists are randomly assigned to each machine and asked to type the same copy matter for 10 minutes. At the end of the period, the words per minute (wpm) are recorded. The data are presented in the table below.

	Output from typewriter (wpm)				
Typewriter	*A*	*B*	*C*	*D*	*E*
A Brand	69	62	70	57	62
B Brand	67	72	76	69	71
C Brand	76	70	71	66	77
D Brand	60	64	67	58	66

Carry out an analysis of variance to assess whether the mean wpm on the different brands of typewriters may be assumed to be the same, or are different.

37. Experiments were performed to determine whether the yield from a chemical process is influenced by the concentration of the catalyst and the temperature of the reaction. Five different concentration levels C_1 to C_5 were combined with three levels of temperature T_1 to T_3.

	Concentration levels of Catalyst				
Temperature Levels	C_1	C_2	C_3	C_4	C_5
T_1	66	72	59	74	68
T_2	64	70	62	73	72
T_3	68	74	64	70	70

Test at 5 per cent significance level, whether the mean yields are influenced by concentration of catalyst or by temperature of reaction.

38. In a certain factory, production can be accomplished by four different types of machines. A sample study, in context of a two-way design without repeated values, is being made with two-fold objectives of examining whether the four workers differ with respect to mean productivity and whether the mean productivity is the same for the five different machines. The researcher involved in this study reports while analysing the gathered data as under:
    (i) Sum of squares for variance between machines = 35.2
    (ii) Sum of squares for variance between workmen = 53.4
    (iii) Sum of squares for total variance = 174.2
    Set up ANOVA table for the given information and draw the inference about variance at 5 per cent level of significance.

39. Three training methods were compared to see if they led to greater productivity after training. Below are productivity measures for individuals trained by each method :

Method 1 :	10	6	8	12	6					
Method 2 :	6	6	7	9	4	6	10	5	6	8
Method 3 :	11	8	13	10	10	12				

At 0.05% level of significance, do the three training methods lead to different levels of productivity ?

40. Perform a two-way ANOVA on the data given below :

Plots of Land	*Treatment*			
	A	*B*	*C*	*D*
I	38	40	41	39
II	45	42	49	36
III	40	38	42	42

41. The following data pertain to the number of units of a product manufactured per day by five workmen from four different brands of machines.

	Machine Brands			
Workmen	*A*	*B*	*C*	*D*
1	46	40	49	38
2	48	42	54	45
3	36	38	46	34
4	35	40	48	35
5	40	44	51	41

    (i) Test, whether the mean productivity is the same for the four brands of machine type.
    (ii) Test, whether five different workmen differ with respect to productivity. *(M.Com., DU,1999)*

42. The following data represent the number of units produced by 4 operators during 3 different shifts :

Shifts	*Operator*			
	A	*B*	*C*	*D*
I	10	8	12	13
II	10	12	14	15
III	12	10	11	14

Perform a two-way analysis of variance and interpret the result. *(MBA, Madras Univ., 2005)*

43. What is 'Analysis of variance' and where it is used ? Given below are the lives (in hours) of three randomly selected batches of electric lamps.

Batch 1	1610	1615	1625	1630	
" 2	1590	1605	1620		
" 3	1580	1585	1600	1610	1625

Analyse the data and draw your conclusions.

For $\alpha = 0.5$, $F_{2,9} = 4.26$, $F_{2,10} = 4.10$, $F_{2,11} = 3.98$
$F_{3,8} = 4.07$, $F_{3,9} = 3.87$, $F_{4,7} = 4.12$ *(M. Com., A.M.U., 2001)*

44. As part of the investigation of the collapse of the roof of a building, a testing laboratory is given all the available bolts that connected the steel structure at three different positions on the roof. The forces required to shear each of these bolts (coded values) are as follows :

Position 1	:	90	82	79	98	83	91	
Position 2	:	105	89	93	104	89	95	86
Position 3	:	83	89	80	94			

Perform an analysis of variance to test at the 0.05 level of significance, whether the differences among the sample means at the three positions are significant. *(B.E./B.Tech, Madras Univ., 2003)*

45. The R & D manager of an automobile company wishes to study the effect of "Tyre Brand" on the tread loss (in millimeters) of tires. Four tyres from each of four different brands (*A, B, C* and *D*) are fitted to four different cars using the completely randomized design. The data as per this design are presented below :

Tyre Brand			
A	*B*	*C*	*D*
6	3	8	4
7	6	6	2
10	2	7	1
9	3	2	4

(*i*) Write the corresponding model.

(*ii*) Check whether the tyre brand has effect on the tread loss of tyres at a significant level of 5%.

(MBA, Bharathidasan Univ., 2002)

46. There are three main brands of a certain powder. A set of 120 sales is examined and found to be allocated among four groups (*A, B, C* and *D*) and brands (*I, II* and *III*) as shown below :

	Replications				
		Groups			
	Brands	*A*	*B*	*C*	*D*
	I	0	4	8	15
Factor	II	5	8	13	6
	III	18	19	11	13

Check whether the factor "Brand" has significant effect on the sales at $\alpha = 0.05$ using one way ANOVA.

(MBA, Bharathidasan Univ., 2006)

47. The following are the number of mistakes made in 5 successive days by 4 technicians working for a photographic laboratory. Test at a level of significance $\alpha = 0.01$, whether the differences among the four sample means can be attributed to chance.

Mistakes	*Technician I*	*Technician II*	*Technician III*	*Technician IV*
Day 1	6	14	10	9
Day 2	14	9	12	12
Day 3	10	12	7	8
Day 4	8	10	15	10
Day 5	11	14	11	11

(MBA, Anna Univ., 2007)

*****

19

Statistical Quality Control

INTRODUCTION

In this era of ever-growing competition, it has become absolutely necessary for businessmen to keep a continuous watch over the quality of the goods produced. Having once bought the product, etc., a kind of goodwill for the product is developed which leads to increase in sales. However, if the consumers are not happy with the quality of product and the complaints are not given proper attention, it shall be impossible for the manufacturer to continue in the market. Either he would improve the quality or else be forced to quit the market by other producers who could start capturing the market by offering better quality.

Although, the need for maintaining and improving quality standard is growing with increasing competition, the idea of quality control is not a new one. What is new about quality control is the use of statistical techniques which are helpful in maintaining and improving quality standards—hence the term **statistical quality control**. Statistical quality control involves the statistical analysis of the inspection data; such analysis is based on sampling and the principles involved in normal curve. These techniques were developed by *W. A. Shewart,* working for the Bell Telephone Co. in the U.S.A.; the problem he solved was that of checking on the consistency of manufacture of a very large number of components. The idea that statistics might be instrumental in controlling the quality of the manufactured products goes back to 1920's and 1930's, but it was not until the pressure of the production needs developed during the period of World War II that its value was fully appreciated. During that period, the use of this technique spread rapidly in British and American war factories and resulted in great savings. For example, in case of Western Electric Company, the rejects of some items declined to 50 per cent which led to a saving of millions of dollars in overheads. In fact, as a result of the war, it became necessary for many different industries to devote themselves to an all-out production effort, requiring the production of tremendous amounts of war material made to more exacting specifications than ever before and in many cases, made by new methods, with substitute materials; poorly trained help with machines designed for other purposes. It was this exigency which led to the wide acceptance of the statistical quality control. Its success in the war was followed by its continued and expanded use in the post-war period. These days, the statistical quality control is used to some extent in virtually every kind of industry in existence. In fact, it has become an integral and permanent part of management controls.

It is important to distinguish between the unsystematic inspection and supervision which often goes under the name of "quality control", and statistical quality control. The former does not say when or how samples should be taken or how large they should be, ordinarily does not have the advantages that go with graphic presentation and does not enforce a clear objective standard for "take action" or "skip it". The statistical quality control chart makes use of well-thought out, tested rules and avoids the indecision, inconsistency and arbitrariness of haphazard quality control. Statistical quality control is based on the fact that repeated random samples from a fixed population will vary, but in a predictable pattern.

The term 'quality' in statistical quality control is usually related to some measurement made on the items produced, a good quality item having one which conforms to standards specified for the measurement. Quality does not always imply the highest standards of manufacture, for the standard required is often deliberately below the highest possible. It is almost always the consistency of manufacturer which represents the most desirable situation rather than the absolute standard which is maintained.

The need for quality control arises because of the fact that even after the quality standards have been specified, some variation in quality is unavoidable. For example, a machine is producing 1,00,000 bolts per day of 2 cm. length. It is very unlikely that all the screws are exactly 2 cm. in length. If the measuring instrument is sufficiently precise we can detect some screws which are slightly less than 2 cm. and some which are slightly more than 2 cm. This leads to a search of the possible causes of variation in the product. The variation of a quality characteristic can be divided under two heads :

(*i*) *Chance variation, i.e.,* variation which results from many minor causes that behave in a random manner and produce slight differences in product characteristic. For example, slight changes in temperature, pressure, metal hardness and similar factors interact randomly to produce slight variations in product quality. This type of variation is permissible, and indeed inevitable, in manufacturing. There is no way in which it can be completely eliminated—when the variability present in a production process is confined to chance variation, the process is said to be in a state of statistical control.

(*ii*) *Assignable variation, i.e.*, those variations that may be attributed to special non-random causes. Such variations can be result of the several factors such as a change in the character of input such as raw material, improper machine setting, broken or worn parts, mechanical faults in plant, adjustment of a machine by an operator, etc.

Out of these two types of variation, nothing can be done about the former type. However, assignable variation can be detected and corrected. The value of quality control lies in the fact that assignable variations in a process can be quickly detected—in fact these variations are often discoverd before the product becomes defective.

There are two different ways of controlling the quality of a product :

(*i*) Through 100% inspection, *i.e.,* by inspecting each and every item, that is produced; and

(*ii*) Through sampling technique or the use of statistical quality control.

The system of 100% inspection is not very satisfactory because of the following reasons :

(*i*) It is too expensive.

(*ii*) It is not always reliable because it becomes too much a routine for the persons inspecting each and every item and defective pieces may also be labelled 'satisfactory'. Defective pieces may also be passed at times when distraction occurs. For example, even when an inspector is trying to perform his task conscientiously, if someone talks to him or someone else happens to attract his attention, he may at times pass faulty pieces.

(*iii*) The inspection is made at the end of the manufacturing cycle, and hence provides few controls over the manufacturing process.

Thus, we find that even 100% inspection is not infallible. In an effort to find out a more economical and yet practical procedure, greater and greater use is being made of the tool of statistical quality control. *Statistical quality control is simply a statistical method for determining the extent to which quality goals are being met without necessarily checking every item produced and for indicating whether or not the variations which occur are exceeding normal expectations.* The statistical control of quality calls for the application of the theory of sampling and tests of significance.

Quality control methods are applied to two distinct phases of plant operation :

(*i*) The control of a **Process** during manufacture. A process is said to be in a state of statistical control, if the variation is such as would occur in random sampling from some stable population. If this

the case, the variation among the items is attributable to chance and there is no point in seeking special causes for individual cases. But when the process is out of control, it should be possible to locate specific causes for the variation and by removing them to improve the future performance of the process. Statistical quality control may be applied to any repetitive process. Such processes are found not merely in machine production in a factory but also in many management problems. Statistical quality control methods have been used in connection with such diverse problems as the stamping out of bottle caps, errors in the work of accountants, the filling of cartons, complaints received from customers, and airline reservations. The statistical tool applied in process control is the **control charts**. The primary objectives of process control are : (*a*) to keep the manufacturing process in control so that the population of defective units is not excessive, and (*b*) assisting in determining whether a state of control exists.

(*ii*) The inspection of materials to determine their acceptability whether they be in raw, semi-finished or completed state. This is known as **acceptance inspection** or **sampling inspection**. The object of acceptance inspection is to evaluate a definite lot of material that is already in existence and about whose quality, a decision must be made. This is done by inspecting a sample of the material, using definite statistical standards to infer from the quality of the sample whether the whole lot is acceptable. The standards in acceptance inspection are set according to what is required of the product rather than by the inherent capabilities of the process, as in process control. In process control, the population is the infinite number of possible results from the same repetitive process. In sampling inspection, the population is the finite group of items which have been produced, usually referred to as a lot.

Control Charts

A control chart is a statistical device principally used for the study and control of repetitive processes.

Walter A. Shewart, its originator, suggested that control chart may serve, firstly, to define the goal or standard for the process that the management might strive to attain; secondly, it may be used as an instrument to attain that goal; and thirdly it may serve as a means of judging whether the goal has been achieved. Thus, it is an instrument to be used in specification, production and inspection.

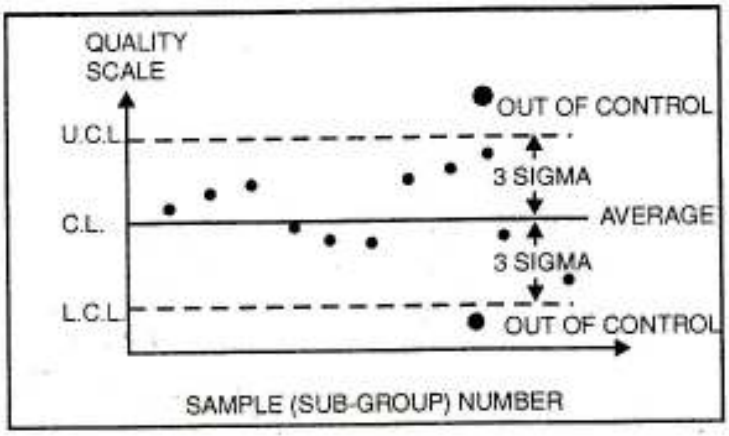
A control chart is essentially a graphic device for presenting data so as to directly reveal the frequency and extent of variations from established standards or goals. Control charts are simple to construct and easy to interpret, and they tell the manager at a glance, whether or not the process is in control *i.e.*, within the tolerance limits. A control chart consists of three horizontal lines :

(*i*) A central line (CL) to indicate the desired standard or level of the process;

(*ii*) Upper control limit (UCL); and

(*iii*) Lower control limit (LCL).

The outline of a control chart is given below :



From time to time, a sample is taken and the data are plotted on the graph. So long as the sample points fall within the upper and lower control limits, there is nothing to worry as in such a case the variation between the samples is attributed to chance or unknown causes.

It is only when a sample point falls outside the control limits that it is considered to be a danger signal indicating that assignable causes are bringing about variations. Thus, there is no wastage of time and money in an effort to find the reason for random variation but as soon as an assignable cause is apparent, necessary corrective action is taken.

Thus, generally, if all dots are found between the upper and lower control limits, it is assumed that the process is "in control" and only chance causes are present. However, sometimes dots are found arranged in some peculiar way. Although they appear between the control limits, a substantial number of successive dots may be located on the same side of the central line or successive dots may follow a definite path leading towards the upper or lower control limit. Such patterns of dots within control limit should also be considered as danger signals which may indicate a change in the production process. Thus, control charts are not only watched for points falling outside the control limits, they are also scrutinised for unusual patterns suggesting trouble.

The control chart may be "likened to a highway whose control limits are the shoulders on one side and centre line on the other. No car driving along the highway can maintain a perfectly straight path. Unevenness in the road, play in the steering, wheel, gusts of wind, and a host of other factors cause slight variations in the movement of the car. It would hardly be worthwhile to investigate the causes of the small irregularities. However, the moment the car moves outside one of the limits, an assignable cause can be assumed to exist and the investigation should begin. The cause may turn out to be a defect in the steering mechanism, a sleepy driver or some similar correctable factor."

How to set up the Control Limits. The basis of control chart is the setting up of upper and lower control limits. These limits are used as a basis for judging the significance of the quality variations from sample to sample, lot to lot or from time to time. The moment a point falls outside these limits, it is taken to be a danger signal. The control limits serve as a guide for action and, therefore, they are also referred to as **action limits**. Control limits are established by computation based upon :

(*i*) Data covering past and current production records ; and

(*ii*) Statistical formulae whose reliability has been proved in practice.

Although, the nature of the control problem does not permit standardising precise and inflexible rules for computing control limits that will be found suited to all of the various conditions that may be encountered in actual practice, it has been found possible to develop certain general procedures on the basis of experience that will cover a wide range of industrial applications.

In most control problems, it had been found statisfactory to place the control limits above and below the grand average of the statistical measures ($\overline{X}$, σ, R, etc.) that is being plotted at distances of three times a computed value, commonly designated as the "sigma" of the statistical measures, for sub-groups of the size under consideration. These are referred to as *"sigma" limits.** The logic of drawing 3σ limits is that in case of a normal distribution, $\overline{X} \pm 3\sigma$ covers 99.73 per cent of the items. In other words, occurrence of events beyond the limits of ($\overline{X} \pm 3\sigma$), provided the events lie on a normal curve, is on the whole nearly 3 out of 1,000 events—an extremely remote chance under normal circumstances. Hence, if points fall outside 3 sigma limits they indicate the presence of some assignable causes—all is not due to random causes. It should be noted that if points fall outside 3 sigma limits, there is a good reason for confidence that they point out to some factor contributing to quality variation that can be identified.

* It should be noted that this value of sigma is not the computed standard deviation of the plotted points. In the case of the $\overline{X}$, R and σ charts, it is computed from the individual observed values with sub-groups and the size n of a sub-group.

The selection of standard value (of $\bar{X}$, σ, R, etc.) is probably the most basic problem encountered in setting up a control procedure. The primary aim is not just to get control, but to get control at a satisfactory level. A satisfactory selection depends fundamentally upon the needs of the buyer or user as defined by his specifications. Any questions of cost of production and capability of manufacturing process must, of course, also be taken into account in deciding on a level that will be economical from an average point of view.

Types of Control Charts

Broadly speaking, control charts can be divided under two heads :

(*i*) Control charts of variables, and

(*ii*) Control charts of attributes.

Variables are those quality characteristics of a product which are measurable and can be expressed in specific units of measurement such as diameter of radio knobs which can be measured and expressed in centimetres, tensile strength of cement which can be expressed in specific measures per square inch of space, etc. Attributes, on the other hand, are those product characteristics which are not amenable to measurement. Such characteristics can only be identified by their presence or absence from the product. For example, we may say that plastic is cracked or not cracked, whether the bottles that have been manufactured contain holes or not. Attributes may be judged either by the proportion of units that are defective or by the number of defects per unit. Thus, the data resulting from inspection of a quality characteristic may take any one of the following forms:

(*i*) A record of the actual measurements of the quality characteristics for individual articles or specimens.

(*ii*) A record of number of articles or specimens inspected and of the number found defective.

(*iii*) A record of the number of defects that are found in a sample. When the possible numbers of defects per sample is very large compared with the average number of defects per sample.

For purposes of control data of the first form, listed above may be summarized by taking two statistical measures, the average ($\bar{X}$) and the standard deviation (σ), or the average ($\bar{X}$) and range (R). Data of the second form can be summarized in terms of fraction defective (p), and third form can be summarized in terms of number of defects per unit.

Setting up a Control Procedure

In establishing basic procedures for the operation of a quality control programme, the manufacturer must take the following preliminary steps :

1. Select the quality characteristics that are to be controlled (including the limits of variation).

2. Analyse the production process to determine the kind and location of probable causes of irregularities.

3. Determine how the inspection data are to be collected and recorded, and how they are to be subdivided.

4. Choose the statistical measures that are to be used in the charts.

Depending on the type of inspection data available, any one of the following types of control charts may be used :

1. *Control charts for* $\bar{X}$ *and* σ; *and* $\bar{X}$ *and* R. Such charts are used when measured values of the quality characteristics are at hand.

2. *Control chart for* $\bar{X}$ *alone.* Control chart for $\bar{X}$ alone is used where experience with control charts for $\bar{X}$ and R, or $\bar{X}$ and σ has demonstrated that instances of lack of control are almost always associated with causes that effect $\bar{X}$ rather than σ or R.

3. *Control chart for* σ *or* R *alone.* Control chart for R or σ is used alone where technical reasons render control of $\bar{X}$ unimportant or where control for $\bar{X}$ is known to be unjustifiably expensive.

4. *Control chart for* C. This chart is used where there are circumstances wherein, the inspection consists of determining the number of defects C in a sample. Such is the case, for example, in the examination of finished textiles, materials, wire, etc.

5. *Control chart for* p *or* pn. Chart for p or pn be used when the records of inspection of testing show merely the number of articles inspected and the number found defective.

These charts are discussed in detail :

$\bar{X}$-Chart. A control chart for sample means, or an $\bar{X}$-chart, is based on the distribution of sample means. It is used to determine if variations in a product dimension are random and to detect assignable variations. The control chart is based on a series of samples or sub-groups of observations drawn randomly from a process over a period of time. The arithmetic means of samples computed and the variation of these means reflect the pattern of variation of the process. The procedure of constructing an $\bar{X}$ - chart is as follows :

1. Obtain the mean of each sample, *i.e.*, $\bar{X}_1, \bar{X}_2, \bar{X}_3$, etc. This is done by dividing the sum of the values included in a sample (ΣX) by the number of observations in the sample (n or sample size).

$$\bar{X} = \frac{\Sigma X}{n}$$

2. Obtain the mean of the sample means, *i.e.*, $\bar{\bar{X}}$. This is done by dividing the sum of the sample means ($\Sigma \bar{X}$) by the number of samples to be included in the chart.

$$\bar{\bar{X}} = \frac{\Sigma \bar{X}}{\text{Number of samples}}$$

3. The control limits are set at

$$UCL = \bar{\bar{X}} + 3\,\sigma_{\bar{X}}$$

$$LCL = \bar{\bar{X}} - 3\,\sigma_{\bar{X}}$$

where, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$; and $\sigma = d/R$

since, R is a biased estimator of σ and d is the correction factor. The values for d are tabulated in the Appendix at the end of the book.

Therefore, the control limits for $\bar{X}$-chart are :

$$UCL = \bar{\bar{X}} + A_2\,\bar{R}$$

$$LCL = \bar{\bar{X}} - A_2\,\bar{R}$$

Illustration 1. A food company puts mango juice into cans advertised as containing 200 grams of the juice. The weights of the juice drained from cans immediately after filling for 20 samples are taken by a random method (at an interval of every [illegible] minutes). Each of the samples includes 4 cans. The samples are tabulated in the following table. The weights in the table are given in units of grams in excess of 200 gms. For example, the weight of a juice drained from the first can of the sample is 215 gms which is in excess of 200 gms excess being 15 gms (215 –200=15). Since the unit in the table is gms, the excess is recorded as 15 units in the table. Construct an $\overline{X}$ chart to control the weights of mango juice for the filling.

Sample number	*Weight of each can (4 cans in each sample n = 4)* X			
1	15	12	13	20
2	10	8	8	14
3	8	15	17	10
4	12	17	11	12
5	18	13	15	4
6	20	16	14	20
7	15	19	23	17
8	13	23	14	16
9	9	8	18	5
10	6	10	24	20
11	5	12	20	15
12	3	15	18	18
13	6	18	12	10
14	12	9	15	18
15	15	15	6	16
16	18	17	8	15
17	13	16	5	4
18	10	20	8	10
19	5	15	10	12
20	6	14	12	14

Solution.

CALCULATIONS FOR CHART

Sample number	*Weight of each can (4 cans in each sample, n = 4)* X				*Total weight of 4 cans* ΣX	*Sample Mean* $\overline{X}$	*Sample Range* R
(1)	(2)				(3)		(4)
1	15	12	13	20	60	15.0	8
2	10	8	8	14	40	10.0	6
3	8	15	17	10	50	12.5	9
4	12	17	11	12	52	13.0	6
5	18	13	15	4	50	12.5	14
6	20	16	14	20	70	17.5	6
7	15	19	23	17	74	18.5	8
8	13	23	14	16	66	16.5	10
9	9	8	18	5	40	10.0	13
10	6	10	24	20	60	15.0	18
11	5	12	20	15	52	13.0	15
12	3	15	18	18	54	13.5	15
13	6	18	12	10	46	11.5	12
14	12	9	15	18	54	13.5	9
15	15	15	6	16	52	13.0	10
16	18	17	8	15	58	14.5	10
17	13	16	5	4	38	9.5	12
18	10	20	8	10	48	12.0	12
19	5	15	10	12	42	10.5	10
20	6	14	12	14	46	11.5	8
Total						263.0	211

Calculations :

(1) The mean of each sample $\bar{X}$ is given in column (3). For example, $\bar{X}$ for the first sample is $\frac{60}{4}$ =15.

(2) The mean of the sample means $\bar{\bar{X}}$ is obtained from column (3) as follows :

$$\bar{\bar{X}} = \frac{\Sigma \bar{X}}{20} = \frac{263}{20} = 13.15$$

(3) The value of R computed from the values of R is shown in column (4). For example, the values of R for the first sample is computed as follows :

$$R = 20 - 12 = 8$$

(4) The value of $\bar{R}$, *i.e.*, the mean of the values of R is obtained as given below :

$$\bar{R} = \frac{\Sigma R}{20} = \frac{211}{20} = 10.55$$

(5) $UCL = \bar{\bar{X}} + A_2 \bar{R}$

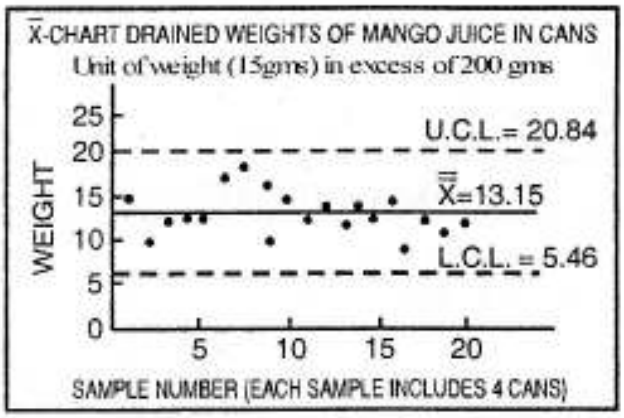
$= 13.15 + 0.729 \times 10.55$ [the table value of A_2 for $n = 4$ is 0.729]

$= 13.15 + 7.69 = 20.84$ approx.

$LCL = \bar{\bar{X}} - A_2 \bar{R}$

$= 13.15 - 0.729 \times 10.55 = 13.15 - 7.69 = 5.46$ approx.

Note that the values in the above computation are expressed in units of gms in excess of 200 gms. The actual values for the *UCL*, thus is 220.208 and that for *LCL* is 205.56 gms. The control chart for this illustration is given below :



Since, all the points are failing within control limits, the process is in a state of control and hence there is nothing to worry.

Illustration 2. A drilling machine bores holes with a mean diameter of 0.5230 cm and a standard deviation of 0.0032 cm. Calculate the 2-sigma and 3-sigma upper and lower control limits for means of samples 4, and prepare a control chart.

Solution. We have

$$\bar{\bar{X}} = 0.5230 \text{ cm}, \sigma = 0.0032 \text{ cm}, n = 4$$

$$\frac{\sigma}{\sqrt{n}} = \frac{0.0032}{2} = 0.0016$$

2- sigma limits for means of sample of 4:

$$UCL = \bar{\bar{X}} + 2\frac{\sigma}{\sqrt{n}}$$

$= 0.5230 + 2(0.0016) = 0.5262$ cm.

Central line = 0.5230 cm.

$LCL = \bar{\bar{X}} - 2(\sigma/\sqrt{n})$

$= 0.5230 - 2(0.0016) = 0.5198$ cm.

3-sigma limits for means of sample of 4 :

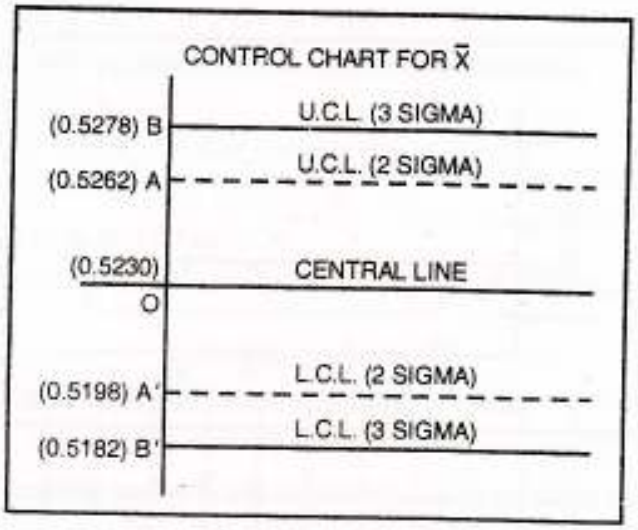
$UCL = \bar{\bar{X}} + 3(\sigma/\sqrt{n})$

$= 0.5230 + 3(0.0016) = 0.5278$ cm.

Central line = 0.5230 cm.

$$LCL = \overline{\overline{X}} - 3(\sigma/\sqrt{n})$$

$$= 0.5230 - 3(0.0016) = 0.5182 \text{ cm.}$$



R-Chart

The *R*-chart is used to show the variability or dispersion of the quality produced by a given process. *R*-chart (or σ chart) is the companion chart to the $\overline{X}$ chart and both are usually required for adequate analysis of the production process under study. The *R*-chart is generally presented along with the $\overline{X}$ chart. The general procedure for constructing the *R*-chart is similar to that of the $\overline{X}$ chart. The required values for constructing the *R*-chart are :

1. The range of each sample, *R*;
2. The mean of the sample ranges, $\overline{R}$;
3. *UCL* and *LCL*

$$UCL_R = \overline{R} + 3\sigma_R, \text{ and}$$

$$LCL_R = \overline{R} - 3\sigma_R$$

where σ_R = The standard error of the range.

The value or σ_R may be estimated by finding the standard deviation of the ranges of the samples included in a chart. In practice, however, it is rather convenient to compute the upper and lower control limits by using the values D_4 and D_2 as provided in Appendix, according to various sample sizes (n = 2 to 20). When the tabulated values are used, the limits may be written as follows :

$$UCL_R = D_4 \overline{R}$$

$$LCL_R = D_3 \overline{R}.$$

It should be noted that the use of *R*-chart is recommended only for relatively small sample size, rarely more than 12 or 15 units. For the large sample sizes ($n > 12$), the σ chart is to be preferred.

Illustration 3. Prepare an *R*-chart for the data of illustration 1.

Solution. The required values for the chart are :

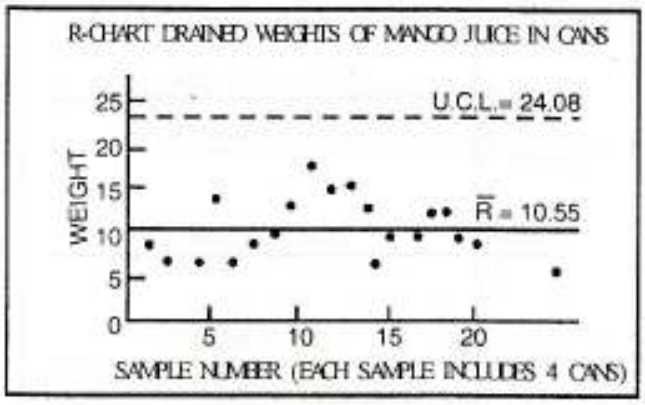
The range of each sample, *R* (see column 4 of illustration 1.).

The mean of the sample ranges, $\overline{R}$

$$\overline{R} = \frac{211}{20} = 10.55$$

$$UCL_R = D_4 \bar{R}$$
$$= 2.282\,(10.55) = 24.08$$ [table value of $D_4 = 2.282$]
$$LCL_R = D_3 \bar{R}$$
$$= 0\,(10.85) = 0.$$

The control chart for R is shown below :



The chart shows that the process is under control since all the R values plotted on the chart are within the two control limits.

The choice between the $\bar{X}$ -chart and the R-chart is a managerial problem.

It is better to construct R-chart first. If the R-chart indicates that the dispersion of the quality by the process is out of control, generally, it is better not to construct an $\bar{X}$ -chart until the quality dispersion is brought under control.

Illustration 4. The table below gives the (coded) measurements obtained in 20 samples (sub-groups). Construct control charts based on the mean and the range. The values of these statistics are given below for the respective samples :

Sub-group

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	−1	2	1	2	1	1	−1	1	2	−2	0	2	0	0	−1	1	2	2	0	3
	2	0	1	1	−1	−1	−1	1	1	1	1	1	1	0	2	−1	1	0	2	−3
	1	1	0	0	0	2	0	2	−1	−2	−3	−1	−3	−1	1	−2	−1	1	1	−1
	0	0	0	−1	0	0	−2	−1	0	2	2	0	2	0	1	0	0	0	−1	1
	1	1	1	0	−1	−2	1	0	0	1	1	0	1	1	2	2	0	1	1	2
$\bar{X}$:	.6	.8	.6	.4	−.2	0	−.6	.6	.4	0	.2	.4	.2	0	1.0	0	.4	.8	.6	.4
R :	3	1	1	3	2	4	3	3	3	4	5	3	5	2	3	4	3	2	3	6

Solution.

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_n}{n_1 + n_2 + n_3 + \ldots}$$
$$= \frac{.6 + .8 + .6 + .4 - .2 + 0 - .6 + .6 + .4 + 0 + .2 + .4 + .2 + 0 + 1.0 + 0 + .4 + .8 + .6 + .4}{20}$$
$$= \frac{6.6}{20} = 0.33$$
$$R = \frac{3+1+1+3+2+4+3+3+3+4+5+3+5+2+3+4+3+2+3+6}{20}$$
$$= \frac{63}{20} = 3.15$$

From the table* for the sample of size 5, we find that

$$A_2 = 0.577,\ D_3 = 0 \text{ and } D_4 = 2.115$$

Upper and Lower control limits for $\bar{X}$ *-chart*

$$= \bar{\bar{X}} \pm A_1\ \bar{R} = 0.33 \pm 0.577\ (3.15) = 0.33 \pm 1.818.$$

Lower control limit = $0.33 - 1.818 = -1.488$.

Upper control limit = $0.33 + 1.818 = 2.148$.

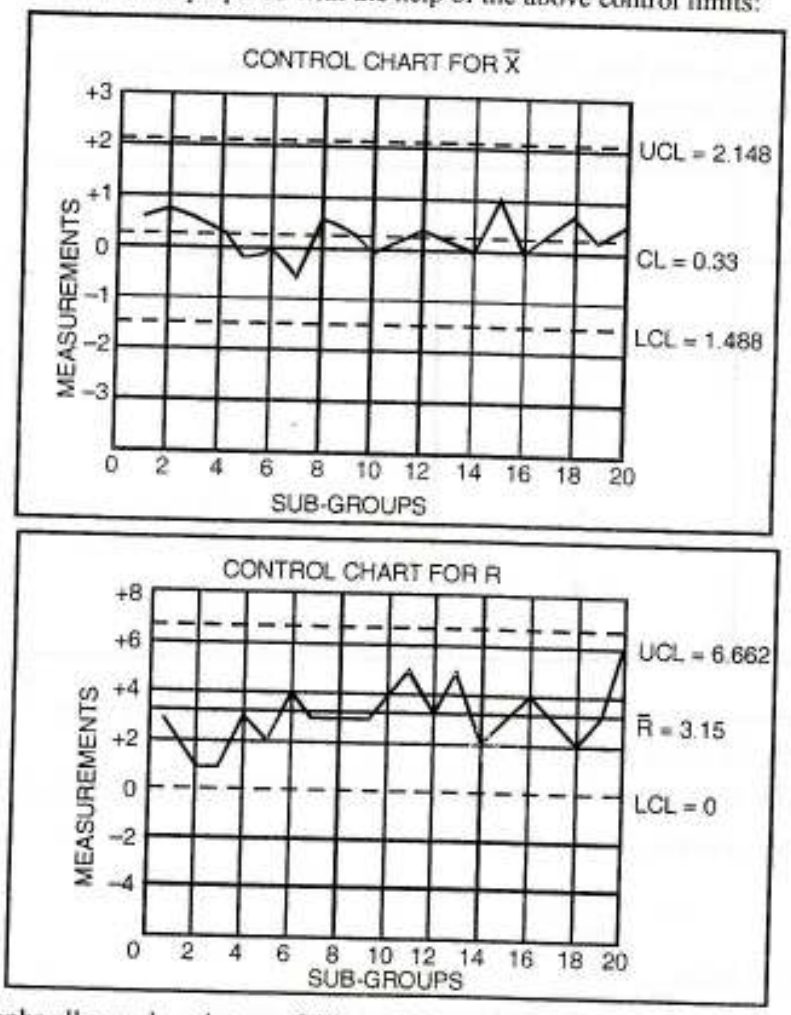
Upper and Lower control limits for R-chart

$$UCL = D_4\ \bar{R} \text{ and } LCL = D_3\ \bar{R}$$

$$UCL = 2.115\ (3.15) = 6.662$$

$$LCL = 0\ (3.15) = 0.$$

The following two control charts are prepared with the help of the above control limits:



The fact that in both graphs all sample points are falling within the 3-sigma control limits, can be interpreted as implying that the process is in a state of statistical control or, in other words, that the only kind of variation present is chance variation.

C-Chart

That C-chart is designed to control the number of defects per unit. The C-chart is based on the Poisson distribution and is very popularly used in statistical work.

Control chart for C is used in situations, wherein the opportunity for defects is large while the actual occurrence tends to be small. Such situations are described by the Poisson distribution. This happens, for example, we count the number of imperfections in a piece of cloth, the number of air bubbles in a piece of glass, the number of blemishes in a sheet of paper, etc. Let C stand for the number

*A_2, D_3 and D_4 are given in the ASTM Manual Table, reproduced and presented at the end of the text. It should be noted that when n is 6 or less, $D_4 = 0$ hence, the lower control limits for R is taken as zero.

of defects counted in one unit of cloth, paper glass, rolls of wire and $\overline{C}$ for the mean of the defects counted in several (usually 25 or more) such units of cloth, the Central Line of the control chart for C is $\overline{C}$ and the 3-sigma control limits are :

$$\overline{C} \pm 3\sqrt{C}.$$

This formula is based on a normal curve approximation of the Poisson distribution. The use of the C-chart is appropriate, if the opportunities for a defect in each production unit are infinite but the probability of a defect at any point is very small and is constant.

Uniform sample size is highly desirable while using the C-chart. Where sample size varies, particularly, if the variation is large, the C-chart becomes difficult to read, and p-chart is the better choice.

Illustration 5. Assume the twenty litre milk bottles are selected at random from a process. The number of air bubbles (defects) observed from the bottles are given in the following table :

[C = No. of Air Bubbles (defects) in each bottle]

Bottle number (Sample order)	*Defects C*	*Bottle number (Sample order)*	*Defects C*
1	4	11	3
2	5	12	5
3	7	13	4
4	3	14	3
5	3	15	4
6	5	16	5
7	6	17	3
8	2	18	7
9	4	19	6
10	8	20	13

Total number of defects = 100

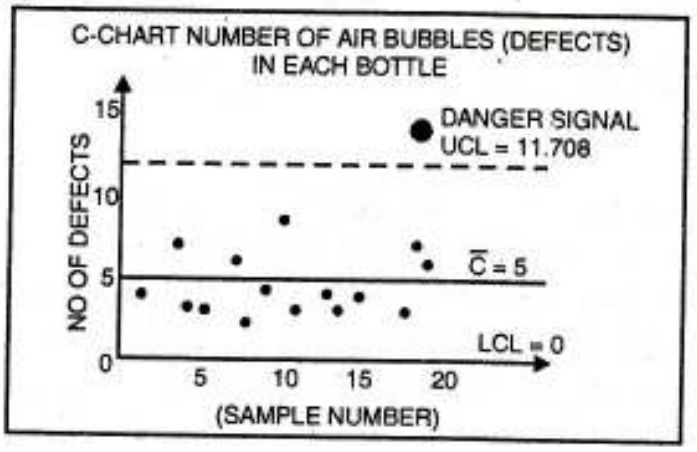
Draw a control chart for the above data.

Solution. We will use the C-chart here. The computation required for preparing this chart are :

$\overline{C}$, *i.e.*, average number of defects

$$\overline{C} = \frac{100}{20} = 5.$$

The control chart is given below :



$$UCL = \overline{C} + 3\sqrt{\overline{C}} = 5 + 3\sqrt{5} = 5 + 3 \times 2.236 = 5 + 6.708 = 11.708.$$

$$LCL = \overline{C} - 3\sqrt{\overline{C}} = 5 - 3\sqrt{5} = 5 - 3 \times 2.236 = 5 - 6.708 = -1.708.$$

The lower control limit will be recorded as zero, since the number of defects cannot be negative.

It is clear from the chart that only one point in respect of last sample falls outside control limits and this is to be treated as danger signal.

Illustration 6. The following table gives the number of errors of alignment observed at final inspection of a certain model of bus. Prepare a C-chart and comment on it.

Bus number	*No. of alignment defects*	*Bus number*	*No. of alignment defects*
1001	6	1011	8
1002	10	1012	6
1003	8	1013	10
1004	7	1014	10
1005	12	1015	6
1006	9	1016	12
1007	5	1017	3
1008	7	1018	11
1009	3	1019	2
1010	4	1020	1

Solution.

$\bar{C}$, *i.e.*, average number of defects $= \frac{140}{20} = 7$.

The control limits and the central line are :

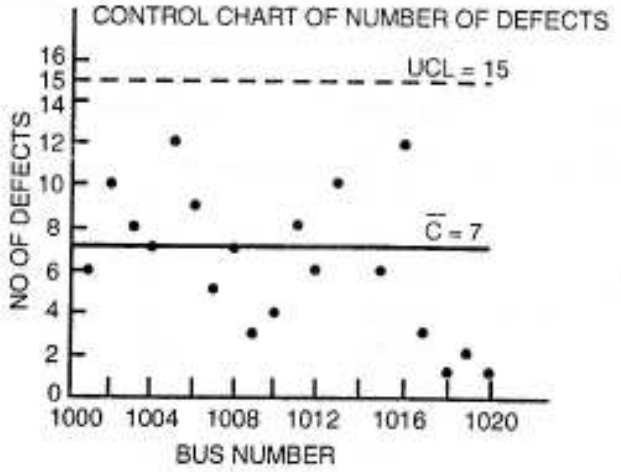
$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$

$$= 7 + 3\sqrt{7} = 7 + (3 \times 2.646) = 14.938 \text{ or } 15.$$

$$\text{Central line} = \bar{C} = 7$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}} = 7 - (3 \times 2.646) = -0.938.$$

Because the number of defects cannot be negative, the lower limit will be taken as zero.



p-Chart

The p-chart is designed to control the percentage or proportion of defectives per sample. A control chart for fraction defective, or p-chart, is based on the distribution of sample proportions. It is assumed that the items are produced by Bernoulli process. This assumption implies that (1) there are only two possible outcomes (acceptable or defective), (2) the outcomes occur randomly, and (3) the probability of either outcome remain unchanged for each trial. Since the number of defectives (c) can be converted into a percentage expressed as a decimal fraction merely by dividing (c) by sample size, the p-chart may be used in lieu of the C-chart. The p-chart has at least two advantages over the C-chart :

The lower control limit will be recorded as zero, since the number of defects cannot be negative.

It is clear from the chart that only one point in respect of last sample falls outside control limits and this is to be treated as [illegible] signal.

Illustration 6. The following table gives the number of errors of alignment observed at final inspection of a certain model [illegible]. Prepare a C-chart and comment on it.

Bus number	*No. of alignment defects*	*Bus number*	*No. of alignment defects*
1001	6	1011	8
1002	10	1012	6
1003	8	1013	10
1004	7	1014	10
1005	12	1015	6
1006	9	1016	12
1007	5	1017	3
1008	7	1018	11
1009	3	1019	2
1010	4	1020	1

Solution.

$\bar{C}$, *i.e.*, average number of defects $= \frac{140}{20} = 7$.

The control limits and the central line are :

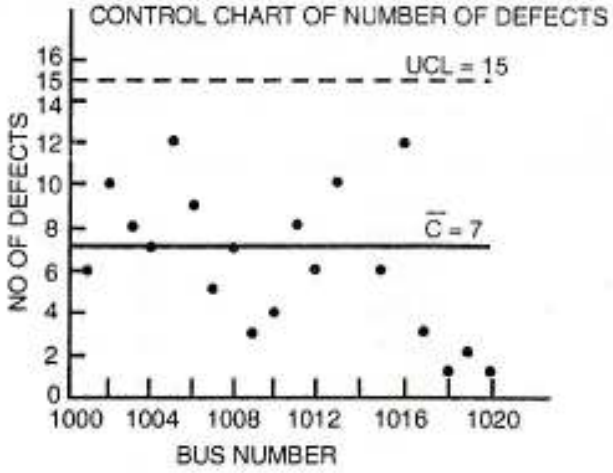
$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$

$$= 7 + 3\sqrt{7} = 7 + (3 \times 2.646) = 14.938 \text{ or } 15.$$

$$\text{Central line} = \bar{C} = 7$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}} = 7 - (3 \times 2.646) = -0.938.$$

Because the number of defects cannot be negative, the lower limit will be taken as zero.



p-Chart

The p-chart is designed to control the percentage or proportion of defectives per sample. A control chart for fraction defective, or p-chart, is based on the distribution of sample proportions. It is assumed that the items are produced by Bernoulli process. This assumption implies that (1) there are only two possible outcomes (acceptable or defective), (2) the outcomes occur randomly, and (3) the probability of either outcome remain unchanged for each trial. Since the number of defectives (c) can be converted into a percentage expressed as a decimal fraction merely by dividing (c) by sample size, the p-chart may be used in lieu of the C-chart. The p-chart has at least two advantages over the C-chart :

1. Expressing the defectives as a percentage or fraction of production is more meaningful and more generally understood than would be the statement of the number of defectives. The latter concept must be related in some way to the total number produced.

2. Where the size of the sample varies from sample to sample, the p-chart permits a more straightforward and less clustered up presentation. The p-chart requires, however, that the division c/n be made. This additional computation may be regarded as a slight disadvantage.

The same basic data is used for either chart. When the sample size remains constant from sample to sample, the primary difference lies in the computation of the control limits. The C-chart control limits are set at $\bar{C}$ plus or minus three standard deviations. The p-chart control limits are set at $\bar{p}$ plus or minus three standard errors of the proportion.

This chart has its theoretical basis in the binomial distribution, and generally give best results when the sample size is large, say, at least 50. The steps in constructing the chart are :

(*i*) Compute the average fraction defective (p) by dividing the number of defective by the total number of units inspected.

(*ii*) On the chart, draw a solid horizontal line to present $\bar{p}$.

(*iii*) Determine the upper and lower control limits. The upper and lower control limits are obtained by the average per cent defective plus and minus three times the standard error as follows :

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \; ; LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

While constructing the chart, it is generally preferred to express results in terms of 'per cent defective' rather than 'fraction defective'. The per cent defective is 100 p. Any sample point falling outside the control limits is evidence of a possible lack of control in as much as the probability of getting such value by chance is less than 0.003. The following example shall illustrate the procedure :

Illustration 7. In a factory producing spark plugs, the number of defectives found in the inspection of 20 lots of 100 each is given as follows :

INSPECTION DATA ON COMPLETED SPARK PLUGS

(2,000 spark plugs in 20 lots of 100 each)

Lot number	*Number defectives*	*Fraction defectives*	*Lot number*	*Number defectives*	*Fraction defectives*
1	5	0.050	11	4	0.040
2	10	0.100	12	7	0.070
3	12	0.120	13	8	0.080
4	8	0.080	14	2	0.020
5	6	0.060	15	3	0.030
6	5	0.050	16	4	0.040
7	6	0.060	17	5	0.050
8	3	0.030	18	8	0.080
9	3	0.030	19	6	0.060
10	5	0.050	20	10	0.100
				Total = 120	

Construct an appropriate control chart.

Solution. Since we are given fraction defective, the suitable chart will be p-chart.

Calculations for p-chart are :

Average fraction defective,

$$\bar{p} = \frac{120}{2{,}000} = 0.06$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

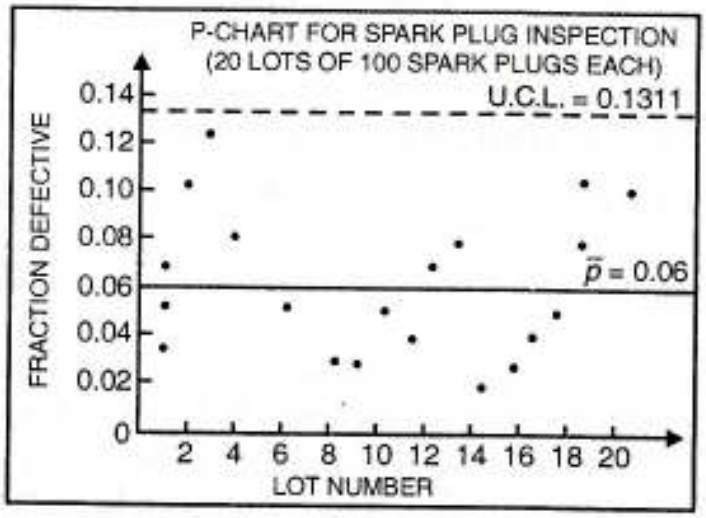
$$= 0.06 + 3\sqrt{\frac{0.06(1-0.06)}{100}} = 0.06 + 3\sqrt{\frac{0.06 \times 0.94}{100}}$$

$$= 0.06 + 3\,(0.0237) = 0.06 + 0.0711 = 0.1311$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 - 3\sqrt{\frac{0.06(1-0.06)}{100}}$$

$$= 0.06 - 0.0711 = -0.0111$$

Since the fraction defective cannot be negative, the LCL shall be taken as zero here.



The control chart shows that all the points are falling within control limits. Hence, the process is in a state of control.

In order to simplify the work of the person who plots the necessary points on the control charts, the above charts can be modified so that he can directly plot the *number* rather than the fraction or percentage of defectives. Such a chart is called the Control Chart for number of defectives. To obtain such a chart, the central line as well as the control limits are multiplied by n. The central line, thus becomes n and the control limits are

$$n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})}$$

Illustration 8. The following data refer to visual defects found at inspection of the first 10 samples of size 100. Use them to obtain upper and lower control limits for percentage defective in samples of 100. Represent the first ten sample results in the chart you prepare to show the central line and control limits.

Sample No.	1	2	3	4	5	6	7	8	9	10	Total
No. of defectives	2	1	1	3	2	3	4	2	2	0	20

Solution. Since there are 20 defective items in 10 samples each of size 100, therefore, $\bar{p}$ = Average fraction defective $= \dfrac{20}{10 \times 100} = 0.02$.

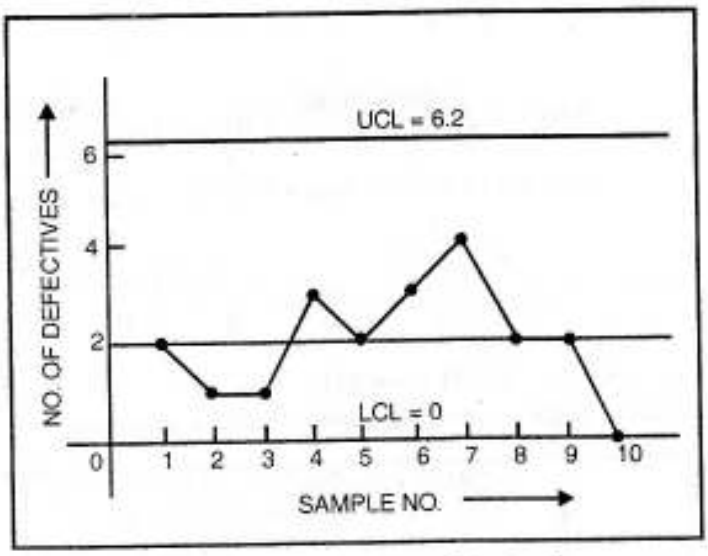
Also, $n = 100$ and $n\bar{p} = 100 \times 0.02 = 2$

and $\sqrt{n\bar{p}(1-\bar{p})} = \sqrt{100 \times 0.02 \times 0.98} = \sqrt{1.96} = 1.4.$

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 2 + 3\,(1.4) = 6.2.$$

Central line $= n\bar{p} = 2.$

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 2 - 3(1.4) = -2.2.$$



Benefits and Limitations of Statistical Quality Control

Advantages. Statistical quality control is one of the tools of scientific management. It has several advantages over 100 per cent inspection. These are:

(*i*) *Reduction in costs*. Since only a fraction of output is inspected, costs of inspection are greatly reduced.

(*ii*) *Greater efficiency*. Not only there is reduction in costs but the efficiency also goes up because much of the boredom is avoided, the work of inspection being considerably reduced.

(*iii*) *Easy to apply.* An excellent feature of quality control is that it is easy to apply. Once the system is established, it can be operated by persons who have not had the extensive specialized training or a highly mathematical background. It may appear difficult, only because the statistical principles on which it is based are unrecognised or unknown. However, as these principles are actually based on commonsense, the quality control method finds wide application.

(*iv*) *Early detection of faults.* Quality control ensures an early detection of faults and hence a minimum waste of reject production. The moment a sample point falls outside the control limits, it is taken to be a danger signal and necessary corrective action is taken. On the other hand, with 100 per cent production, unwanted variations in quality may be detected at a stage, when a large amount of faulty products have already been produced. Thus, there would be a big wastage. Control chart, on the other hand, provides a graphic picture of how the production is proceeding and to tell management where not to look for trouble.

(*v*) *Adherence to Specifications.* Quality control enables a process to be brought into and held in a state of statistical control, *i.e.*, a state in which variability is the result of chance causes alone. So long as a statistical control continues, specifications can be accurately predicted for the future, which even 100 per cent inspection cannot guarantee. Consequently, it is possible to assess whether the production processes are capable of turning out products which will comply with the given set of specifications.

(*vi*) *The only course.* In certain cases, 100% inspection cannot be carried out without destroying all the products inspected : for example, testing breaking strength of chalks, proofing of ammunition, etc. In such cases, if 100% inspection methods are followed then all the items inspected will be spoiled. In such a case, sampling must be resorted to and with the application of SQC techniques not only that the quality is controlled but also that valid inferences about the total output are drawn from the samples.

(*vii*) *To determine effect of changed process.* With the help of control charts one can easily detect whether or not a change in the production process results is a significant change in quality.

(*viii*) *Statistical quality control ensures overall coordination.* Statistical quality control provides a basis upon which the difference arising among the various interests in an organisation can be resolved. In some instances, for example, production engineers may set specifications that are so "tight" that the operating staff cannot meet them economically and consequently there is an unnecessary high scrapping rate. In other instances, the specifications may be too loose, and product quality will be sacrificed unnecessarily. In either type of case, the control records provide a valuable aid in solving the problem of getting the operating and engineering forces together on the basis of common understanding. Information on plant capabilities and customer requirements must also be considered in relation to the quality control limits and records of performance and, finally, it should be possible to determine the best practical balance between the cost of quality and the sales value of the product.

SQC has a special role to play in a country like India because of the extraordinary variations encountered in raw materials and in machines. The importance of applying SQC has become greater in our industries in the context of the need for earning foreign exchange by supplying quality goods to successfully compete in the world markets.

Limitations

Despite the great significance of statistical quality control, it should be remembered that it is not a panacea for all quality ills. The techniques of quality control should not be used mechanically rather they should be matched to the process being studied. The application of standard process without adequate study of the process is extremely dangerous, and has in the past led to statistical methods being discredited. Statistical methods applied on a production process are only an information service, and as such must be conditioned by the process to which they are applied. Unless they are used as part of a generally quality awareness, they may only lead to a false sense of security. The responsibility for quality and process decisions rests with the manager incharge of the process and not with the statistician. The charts do not reduce the manager's responsibility.

ACCEPTANCE SAMPLING

The control charts described above cannot be applied to all types of problems. They are useful only to the regulation of the manufacturing process. Another important field of quality control is acceptance sampling. Inspection for acceptance purposes is carried out at many stages in manufacturing. For example, there may be inspection on incoming materials and process inspection at various points in the manufacturing operations, final inspection by a manufacturer of his own product, and ultimately, inspection of the finished product by one or more purchasers. Much of the acceptance inspection is carried out on a sampling basis. The use of sampling inspection by a purchaser to decide whether or not to accept a shipment of product is known as *acceptance sampling.* A sample of the shipment is inspected and if the number of defective items is not more than a stated number known as the *acceptance number,* the shipment is accepted. The standards in acceptance inspection are set according to what is required of the product, rather than by the inherent capabilities of the process, as in the process control. The purpose of acceptance sampling is, therefore, whether to accept or reject a product—it does not attempt to control quality during the manufacturing process, as do the techniques described earlier in the chapter. Sampling inspection may also be referred to as *product control,* because it is designed to provide decision procedures under which, a lot will be accepted or rejected.

Acceptance sampling procedures which were perfected during World War II to meet military needs for quick and accurate inspection of vast supplies of material are now used widely in industry.

A typical application of acceptance is to determine whether a batch of items, called an *inspection lot* or simply a lot, that has been delivered by a supplier, is of acceptable quality. Another application is

to a lot that is complete and ready for shipment to customers to make sure that it is of adequate quality. Still another application is to a lot of partly completed material, to determine whether it is of high quality to justify further processing.

Role of Acceptance Sampling

Acceptance sampling is very widely used in practice because of the following reasons :

1. Acceptance sampling is much less expensive than 100 per cent inspection.

2. In many cases it provides better outgoing quality. It is generally agreed that good 100 per cent inspection will remove only about 85 to 95 per cent of the defective material. Very good 100 per cent inspection will remove 99 per cent of the defective items but still not reach 100 per cent. Because of the effect of inspection fatigue involved in 100 per cent inspection, a *good* sampling plan may actually give better quality assurance than 100 per cent inspection. The word '*good*' is *italicised* since many informal sampling plans devised without benefit of knowledge of the laws of change are practically worthless. The result has been widespread use of sampling plans.

3. In modern manufacturing plants, acceptance sampling is used for evaluating the acceptability of incoming lots of raw materials and parts at various stages of manufacture, and final inspection of finished product.

4. Where quality can be tested only by destroying items, as in determining the strength of glass containers, 100 per cent inspection is out of the question and sampling must be used. Of course, there are situations where 100 per cent inspection is not to be put aside ; for example, in testing rifles to be used by soldiers, we cannot risk imperfection in any item and therefore must test each and every rifle.

Since under a sampling inspection plan, a decision is made as to whether to accept a lot or reject a lot on the basis of sample, there is a possibility (1) rejecting a lot as unsatisfactory when it is of acceptable quality, and (2) accepting a lot as satisfactory when in fact it is below the quality level. Hence in any acceptance sampling plan, the producers and the consumers, the sellers and the buyers, are exposed to some risks. These are called *producer's* and *consumer's risks*. The producer's risk is the risk a producer takes that a lot will be rejected by a sampling plan even though it conforms to requirements. This is equivalent to the concept of type I error, or the probability of rejecting a hypothesis when it is in fact true. The consumer's risk is the risk that a lot of certain quality will be accepted by a sampling plan. It is equivalent to type II error which is the probability of accepting a hypothesis when an alternative is true. Before agreeing to an acceptance criterion, the consumers and producers will like to know the risks to which they are exposed, *i.e.*, the probability of rejecting a good lot and the probability of accepting a bad one.

An inspection plan can easily be constructed if the consumers and producers specify these probabilities and also the proportion of defectives above which a lot is considered to be bad and the proportion of defectives below which a lot is considered to be good.

Types of Acceptance Sampling Plans

The following three types of acceptance sampling plans are commonly used :

1. Single Sampling Plan. When the decision whether to accept a lot or reject a lot is made on the basis of only one sample, the acceptance plan is described as single sampling plan. This is the simplest type of sampling plan. In any systematic plan for single sampling three things are specified, namely, (*a*) Number of items N in the lot from which the sample is to be drawn, (*b*) Number of articles n in the random sample drawn from the lot, and (*c*) The acceptance number c. This acceptance number is the maximum allowable number of defective articles in the sample. More than this will cause the rejection of the lot. Thus, a sampling plan may be specified in this way—

$$N = 200, n = 20, c = 1.$$

These numbers may be interpreted as saying, "Take a random sample of 20 from a lot of 200. If the sample contains more than 1 defective, reject the lot ; otherwise accept the lot."

2. Double Sampling Plan. In the single sampling plan discussed above, decision with regard to acceptance or rejection of a lot is based on the evidence of only one sample from the lot. However, double sampling involves the possibility of putting off the decision on the lot until a second sample has been taken. A lot may be accepted at once if the first sample is good enough or rejected at once if the first sample is bad enough. If the first sample is neither good enough nor bad enough, the decision is based on the evidence of the first and second sample combined. In a double sampling plan, 5 things are specified : n_1, c_1, n_2, $n_1 + n_2$ and c_2, where

n_1 = number of pieces in the first sample;

c_1 = acceptance number for the first sample, the maximum number of defectives that will permit the acceptance of the lot on the basis of the first sample;

n_2 = number of pieces in the second sample;

$n_1 + n_2$ = number of pieces in the two samples combined; and

c_2 = acceptance number for the two samples combined, the maximum number of defectives that will permit the acceptance of the lot on the basis of the two samples.

Thus, a double sampling plan may be :

$$N = 500, n_1 = 20, c_1 = 1, n_2 = 60, c_2 = 4.$$

This will be interpreted as follows :

(*i*) Inspect a first sample of 20 from a lot of 500.

(*ii*) Accept the lot on the basis of the first sample, if it contains 1 defective.

(*iii*) Reject the lot on the basis of the first sample, if the sample contains more than 1 defective.

(*iv*) Inspect a second sample of 60, if the first sample contains 2, 3, 4 defectives.

(*v*) Accept the lot on the basis of combined sample of 80, if the combined sample contains 4 or less defectives.

(*vi*) Reject the lot on the basis of combined sample, if the combined sample contains more than 4 defectives.

Advantages of Double Sampling Plan

A double sampling plan has two possible advantages over a single sampling plan :

(*i*) It may reduce the total amount of inspection ; for the first sample taken is less than that called for under a comparable single sampling plan, and, consequently, in all cases in which a lot is accepted or rejected on the first sample, there may be considerable saving in total inspection. It is also possible to reject a lot without completely inspecting the entire sample.

(*ii*) A double sampling plan has the psychological advantage of giving a lot a second chance. To some people, especially the producer, it may seem unfair to reject a lot on the basis of a single sample. Double sampling permits the taking of two samples on which to make a decision.

3. Multiple or Sequential Sampling Plan. Just as double sampling plans may defer the decision on acceptance or rejection until a second sample has been taken, other plans may permit any number of samples before a decision is reached. Plans permitting from three up to an unlimited number of samples are described as multiple or sequential. However, such plans are quite complicated and rarely used in practice.

Selection of a Sampling Plan

All practical sampling plans have an OC curve. The following points, need emphasis regarding the OC curve :

1. There is some chance that good lots will be rejected.
2. There is some chance that bad lots will be accepted.
3. These risks can be calculated by the theory of probability and depend on the number of samples inspected, the acceptance number and the per cent defective in the lots submitted for sample inspection. Given the amount of risks which can be tolerated, a sampling plan can be derived to meet these requirements.
4. The larger the sample used in sample inspection, the nearer the OC curve approaches the ideal. However, beyond a certain point, the added cost in inspecting a large number of parts far exceeds the benefits derived.

In review, the two parameters of an OC curve are the sample size and the acceptance number. The desired quality level (p) and the probability of acceptance (P_a) must be selected so that the proper sampling plan can be designed.

There are four factors which should be decided in a sampling plan :

1. P_1, also known as AQL (the Acceptable Quality Level). This is the definition of a good lot.
2. P_2, also known as RQL (the Rejectable Quality Level) or $LTPD$ (Lot Tolerance Per cent Defective).
3. α, also known as Producer's Risk. This is the probability of rejecting a good lot.
4. β, also known as the Consumer's Risk. This is the probability of accepting a poor lot.

Construction of an OC Curve

An OC curve can be constructed by using either the Poisson distribution of the Thorndike chart. The Poisson distribution can be used in all situations where p is less the 0.10 (or if the pn is less than 5) and the lot size is at least 10 times the size of the sample.

In a situation in which these conditions are not met, the theoretically correct approach is to use the binomial or the hypergeometric distribution can be used without serious loss of accuracy.

To use the Thorndike chart, the following procedure is followed. For each possible value of the lot, fraction defective p, a pn is computed. The Thorndike chart is used to find the probability of C or less defective units. For example, for a lot that is 5 per cent defective ($p = 0.05$) and a sample size of 100 ($n = 100$), *i.e.*, $pn = 5$, the probability of selecting 2 or less defectives is found from the Thorndike chart to be approximately 0.12. If the lot fraction defective is 1 per cent ($p = 0.01$) and the sample size is 100 ($n = 100$), *i.e.*, $pn = 1$, the probability of selecting 2 or less defective is found from the Throndike chart to be approximately 0.92. These results give two plots on the OC curve for the sampling inspection plan where the sample size is 100 and the acceptance number is 2. Other points may be calculated in the same way.

The Operating Characteristic (OC) Curve

In judging various acceptance sampling plans, it is desirable to compare their performance over a range of possible quality levels of submitted product. An excellent picture of this performance is given by the Operating Characteristic curve. Such curves are commonly referred to as OC curves. The OC curve of an acceptance sampling plan shows the ability of the plan to distinguish between good and bad lots. For any given fraction defective p in a submitted lot, the OC curve shows the probability p_a that such a lot will be accepted by the given sampling plan or, in other words, the OC curve shows the long-run percentage of submitted lots that would be accepted if a great many lot of any stated quality was submitted for inspection. In drawing the OC curve, the following two terms are important.

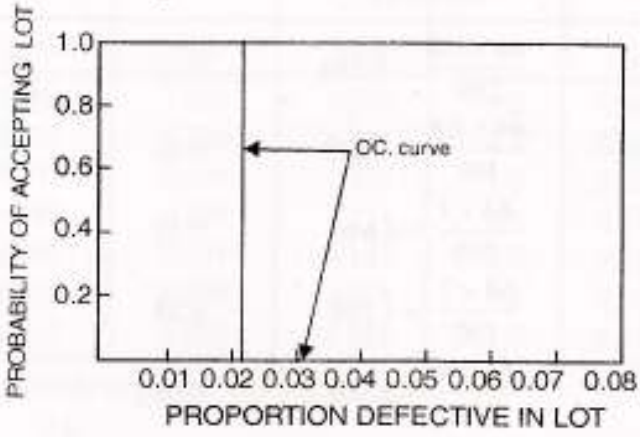
AQL and LTPD

In order to measure the customer's risk, we must define maximum percentage of defective items in lots which the consumer wishes to accept. This is called the *Lot Tolerance Percentage*

Defective or LTPD. Similarly, to measure the producer's risk we define a minimum percentage of defective items in a lot below which the lot should be accepted—this is known as *Acceptable Quality Level* or AQL. The producer's risk is now defined as the probability that a lot having the AQL will be rejected and the consumer's risk as the probability that a lot having LTPD will be accepted. These risks are usually taken as 5% and 10% respectively. The actual levels of the AQL and LTPD must be decided by negotiations between the consumer and the producer.

Shape of an Ideal OC Curve

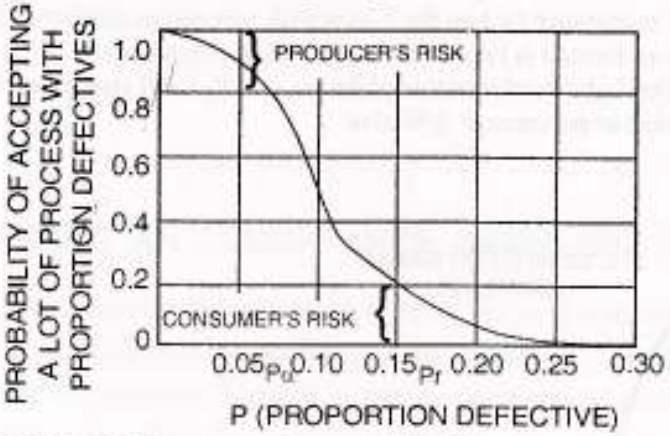
The ideal OC curve would be one for which all good lots are accepted and all bad lots are rejected. Such a curve would look like the following :



No sampling plan can have an *OC* curve of this type. The degree to which an actual *OC* curve approximates the ideal curve depends upon n and c, n representing the sample size and c the acceptance number, or the number of defects in the sample which is not to be exceeded.

Shape of a Typical OC Curve

A typical *OC* curve resembles the following diagram :



The points on the horizontal scale represent possible lot or process qualities, and the height of the curve shows the probability that a lot of this quality will be accepted, assuming the specified sampling plan is in use. In the above diagram, it has been assumed that the acceptable and rejectable qualities are measured as proportions of the items that are defective and are $P_a = 0.05$ and are $P_r = 0.15$; from the *OC* curve, the producer's and consumer's risks are seen to be both a little more than 0.10 in this example. (The sampling plan of the above diagram calls for accepting the lot if three or fewer defectives are found in a sample of 40).

The steepness of the *OC* curve depends upon the sample size. The larger the sample, the steeper the curve, and the smaller the zone between the qualities that are almost always accepted and the qualities that are almost always rejected.

The location of the *OC* curve is determined by the maximum number of defective items allowable for acceptance, called the *acceptance number*. If the acceptance number is made large, the curve is shifted to the right. If the acceptance number is made smaller, the curve is shifted to the left.

Illustration 9. For the sampling plan $N = 1{,}200$, $n = 64$ and $C = 1$ determine the probability of acceptance of the following lots: (*i*) 0.5% defective, (*ii*) 0.8% defective, (*iii*) 1% defective, (*iv*) 2% defective, (*v*) 4% defective, (*vi*) 10% defective. Also draw an *OC* curve.

Solution. If the lot contains 0.5% defective, the samples from it will also have an average of 0.5% defective. Hence in a sample of size 64, the average number of defective will be $\frac{64 \times 0.5}{100} = 0.32$. If the sample contains 1 or 0 defective, the lot is to be accepted under the sampling plan. We can obtain the cumulative probability on drawing a sample of 64 containing 0 or 1 defective by using the Poisson approximation to the binomial distribution. The calculation will be as follows :

S. No.	*% defective in the lot*	*Average number of defectives*	*P(0)*	*P(1)*	*P(0) + P(1) P(a)*
(*a*)	0.5	$\frac{64 \times 0.5}{100} = 0.320$	0.73	0.23	0.96
(*b*)	0.8	$\frac{64 \times 0.8}{100} = 0.512$	0.60	0.31	0.91
(*c*)	1	$\frac{64 \times 1}{100} = 0.640$	0.53	0.35	0.88
(*d*)	2	$\frac{64 \times 2}{100} = 1.280$	0.28	0.36	0.64
(*e*)	4	$\frac{64 \times 4}{100} = 2.560$	0.08	0.21	0.29
(*f*)	10	$\frac{64 \times 10}{100} = 6.400$	0.002	0.01	0.01

The value 0.96 represents the probability of drawing a sample of 64 with 0 or 1 defective from a lot known to be 0.5% defective. Conversely, we can state that such a sample will enable acceptance of 96 per cent of lots containing 0.5 per cent defectives. In other words, if 1,000 such lots are submitted for inspection under the sampling plan, on an average 960 lots will be accepted and 40 will be rejected.

If we take the probabilities of acceptance (p_a) on the *Y*-axis with percentage defective in the lots submitted on the *X*-axis, and join the various points, the curve so formed is known as the operating characteristic (*OC*) curve of the sampling plan. From the *OC* curve, we can easily obtain the probability of rejection of the lot, ($1 - P_a$) will give the probability of rejection corresponding to any lot, having a specified proportion or percentage defective.

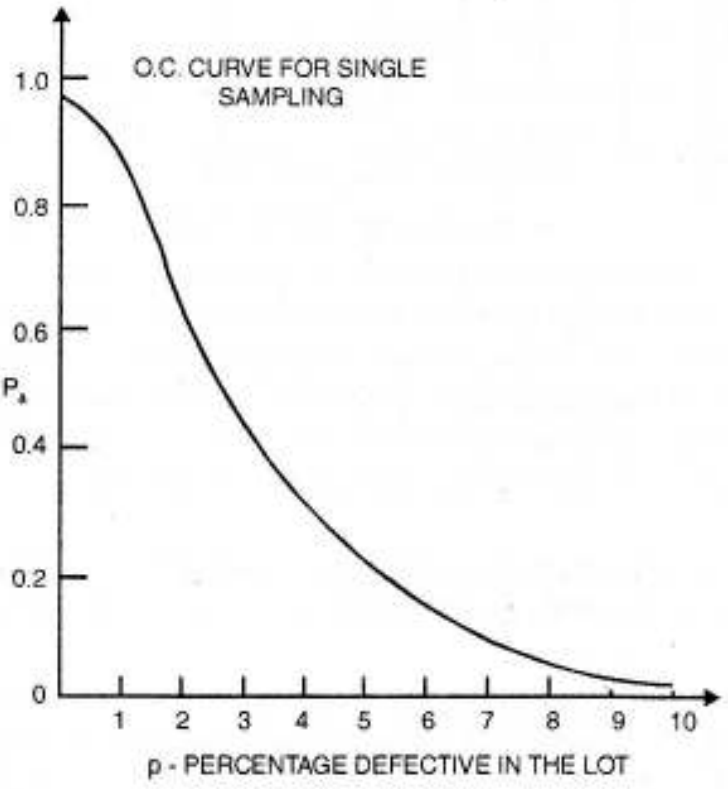


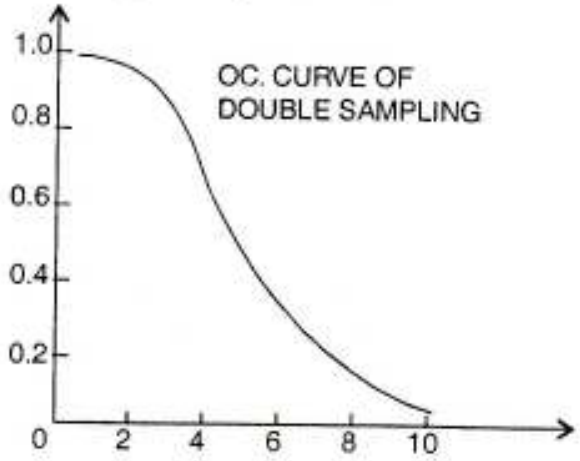
Illustration 10. Draw the *OC* curve of the double sampling plan, given that $N = 1{,}000$, $n_1 = 50$, $c_1 = 1$, $n_2 = 25$, $c_2 = 5$.

Solution. This sampling plan means that a sample of size 50 is drawn, if it gives 0 or 1 defective it is accepted. If it gives 3 or more defective, it is rejected. But, if it gives 2 defectives, then a sample of 25 is drawn. If the total number of defectives is 2, the lot is accepted ; if it is more than 2, the lot is rejected. For various values of percentage defective in lot x, the probability of acceptance shall be obtained as follows :

$$m = \frac{x \times n}{100}, p(0) = e^{-m}, p(1) = e^{-m} \times m, \text{ etc.}$$

1st sample $n_1 = 50$						*2nd sample* $n_2 = 25$			*Combined Sample*
x	m	$p(0)$	$p(1)$	$p(a)$	$p(2)$	m	$p(0)$	$p(a)$ $p(2) \times p(0)$	$p(a)$
0	0	1.000	0.000	1.000	0.000	0.0	0.000	0.000	1.000
2	1	0.368	0.368	0.736	0.184	0.5	0.606	0.112	0.848
4	2	0.135	0.271	0.409	0.271	1.0	0.368	0.100	0.506
6	3	0.050	0.149	0.199	0.224	1.5	0.273	0.049	0.248
8	4	0.018	0.073	0.091	0.147	2.0	0.135	0.020	0.111
10	5	0.007	0.033	0.040	0.084	2.5	0.082	0.007	0.047
12	6	0.002	0.015	0.017	0.045	3.0	0.050	0.002	0.019

Let us plot these points on the graph paper to get the required *OC* curve.



EVALUATING AN ACCEPTANCE SAMPLING PLAN

An acceptance sampling plan can be evaluated in several ways. First, we may ask how the sampling plan discriminates between lots of different incoming quality. Secondly, we are interested in the average outgoing quality of the lots after inspection. Finally, a major concern in evaluating a plan is cost. What shall be the inspection costs under the plan ?

MISCELLANEOUS ILLUSTRATIONS

Illustration 11. Samples of 100 tubes are drawn randomly from the output of a process that produces several thousand units daily. Sample items are inspected for quality and defective tubes are rejected. The result of a series of 20 samples is shown below :

Sample No.	*No. Inspected*	*No. Defectives*	*Sample No.*	*No. Inspected*	*No. Defectives*
1	100	8	11	100	17
2	100	10	12	100	14
3	100	12	13	100	13
4	100	8	14	100	15
5	100	7	15	100	8
6	100	11	16	100	6
7	100	13	17	100	10
8	100	5	18	100	7
9	100	10	19	100	4
10	100	12	20	100	10

Set up the upper and lower control limits.

Solution.

$$\bar{p} = \frac{200}{2000} = 0.1$$

$$\text{UCL} = 0.1 + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.1 + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{100}} = 0.1 + 3(0.03) = 0.19$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.1 - 3(0.03) = 0.01$$

Illustration 12. A random sample of 200 was taken from daily production of large output of pens and number of defective pens was noted. On the basis of information given below, prepare a control chart for fraction defective. What conclusion do you draw from the control chart ?

Production each day	*No. of defectives*	*Production each day*	*No. of defectives*
1	10	13	8
2	5	14	14
3	10	15	4
4	12	16	10
5	11	17	12
6	9	18	11
7	22	19	26
8	4	20	13
9	12	21	10
10	24	22	9
11	21	23	11
12	15	24	12

Solution.

Production each day	*No. of defectives*	*Fraction defectives*	*Production each day*	*No. of defectives*	*Fraction defectives*
1	10	0.050	13	8	0.040
2	5	0.025	14	14	0.070
3	10	0.050	15	4	0.020
4	12	0.060	16	10	0.050
5	11	0.055	17	12	0.055
6	9	0.045	18	11	0.055
7	22	0.110	19	26	0.130
8	4	0.020	20	13	0.065
9	12	0.060	21	10	0.050
10	24	0.120	22	9	0.045
11	21	0.105	23	11	0.055
12	15	0.075	24	12	0.060

$$\bar{p} = \frac{\text{Total number of defectives}}{\text{Total production}} = \frac{294}{24 \times 200} = 0.061$$

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.01 + 3\sqrt{\frac{0.061\,(1-0.061)}{200}} = 0.061 + 3\,(0.0169) = 0.112$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.061 - 3\,(0.0169) = 0.0103.$$

Since all the points don't fall within control limits, there seems to be something wrong with the production process.

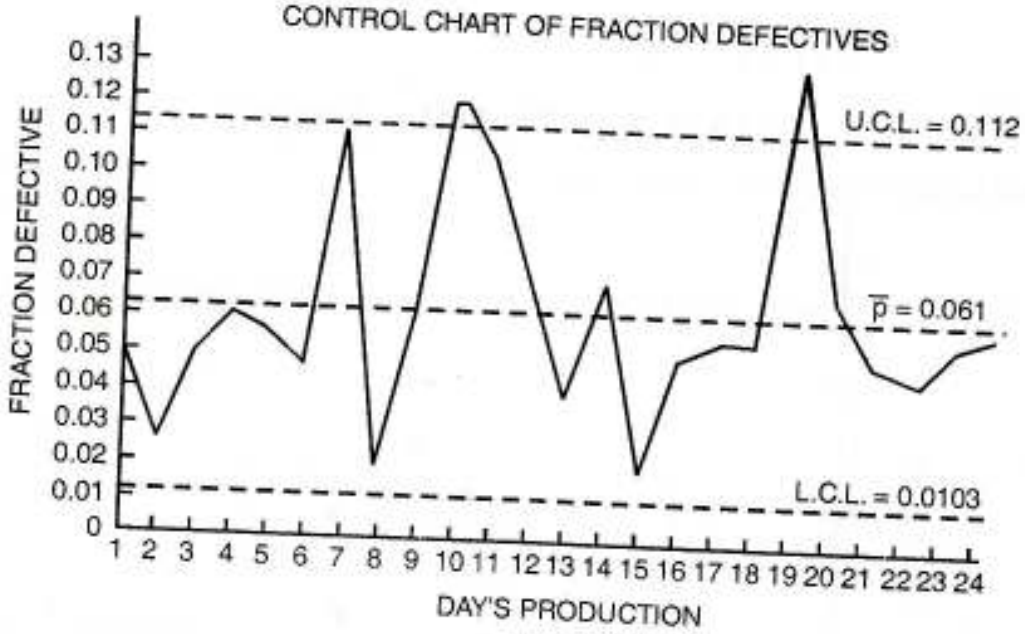


Illustration 13. 20 Television sets were examined for quality control test. The number of defects for each television set are recorded below :

2, 4, 3, 1, 1, 2, 5, 3, 6, 7, 3, 1, 4, 2, 1, 3, 4, 6, 1, 1.

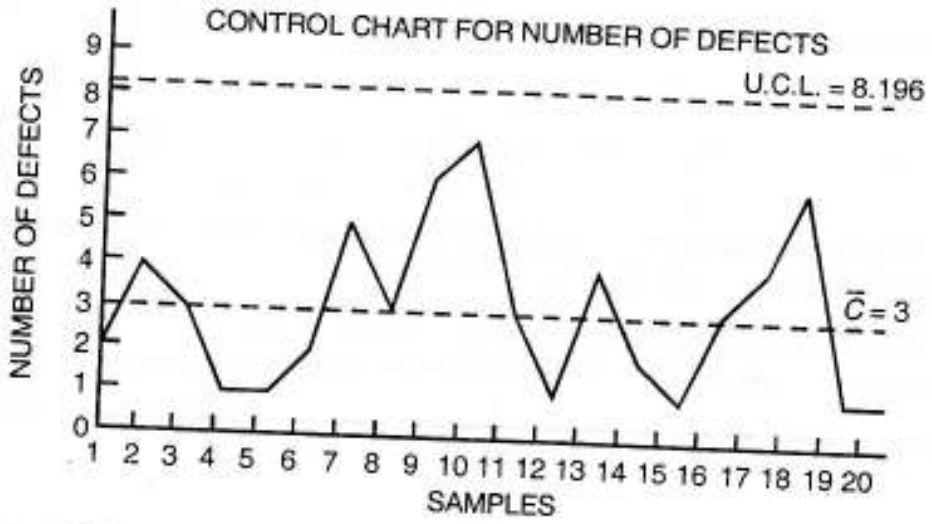
Prepare a C-chart. What conclusion can you draw from it ?

Solution. Total number of defects = 60, Sample size = 20

$$\overline{C} = \frac{60}{20} = 3$$

$$\text{UCL} = \overline{C} + 3 = 3 + 3\sqrt{3} = 3 + 5.196 = 8.196$$

$$\text{LCL} = \sqrt{C} - 3 = 3 - 3\sqrt{3} = 3 - 5.196 = -2.196 \text{ or } 0$$



Since all the points are lying within control limits, the process is in a state of control.

Illustration 14. During an examination of equal length of cloth, the following number of defects are observed :

2, 3, 4, 0, 5, 6, 7, 4, 3, 2.

Draw a control chart for the number of defects and comment whether the process is under control or not.

Solution. Let C denote the number of defects per piece.

$$\Sigma C = 2 + 3 + 4 + 0 + 5 + 6 + 7 + 4 + 3 + 2 = 36.$$

$$\overline{C} = \frac{36}{10} = 3.6$$

Hence central line = 3.6

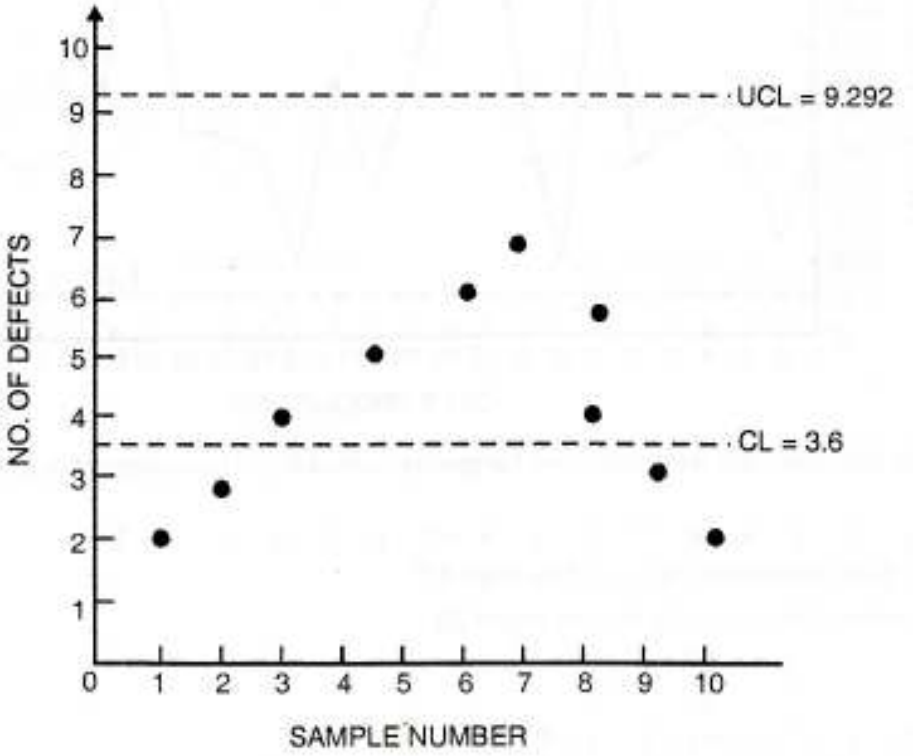
$$\text{UCL} = \overline{C} + 3\sqrt{\overline{C}}$$

$$= 3.6 + 3\sqrt{3.6} = 3.6 + 5.692 = 9.292$$

$$\text{LCL} = \bar{C} - 3\sqrt{\bar{C}}$$

$$= 3.6 - 3\sqrt{3.6} = 3.6 - 5.692 = -2.092 \text{ or } 0$$

The control chart based on these limits is given below :



Since all the points are lying within control limits, the process is in a state of control.

Illustration 15. A machine is set to deliver packets of a given weight. 10 samples of size 5 each were recorded. Below are given relevant data :

Sample No. :	1	2	3	4	5	6	7	8	9	10
Mean ($\bar{X}$) :	15	17	15	18	17	14	18	15	17	16
Range (R) :	7	7	4	9	8	7	12	4	11	5

Calculate the values for the central line and the control limits for mean chart and then comment on the state of control. (Conversion Factors for $n = 5$, are $A_2 = 0.58$, $D_3 = 0$, $D_4 = 2.11$.)

Solution. CALCULATION OF CONTROL LIMITS FOR $\bar{X}$ AND R CHART

Sample No.	$\bar{X}$	R
1	15	7
2	17	7
3	15	4
4	18	9
5	17	8
6	14	7
7	18	12
8	15	4
9	17	11
10	16	5
$n = 10$	$\Sigma\bar{X} = 162$	$\Sigma R = 74$

Control limits for $\bar{X}$ chart

$$\text{UCL} = \bar{\bar{X}} + A_2\bar{R}$$

$$\text{LCL} = \bar{\bar{X}} - A_2\bar{R}$$

$$\text{Central line} = \bar{\bar{X}} = \frac{\Sigma \bar{X}}{n} = \frac{162}{10} = 16.2$$

$$\text{UCL} = 16.2 + 0.58\,(7.4) = 16.2 + 4.292 = 20.492$$
$$\text{LCL} = 16.2 - 0.58\,(7.4) = 16.2 - 4.292 = 11.908$$

Control limits for R chart

$$\text{Central line} = \bar{\bar{X}} = \frac{74}{10} = 7.4$$

$$\text{UCL} = D_4 \bar{R}$$
$$\text{LCL} = D_3 \bar{R}$$
$$\text{UCL} = 2.11(7.4) = 15.614$$
$$\text{LCL} = 0(7.4) = 0$$

Since all the points are lying within control limits, the process is in a state of control.

Illustration 16. Compute the values for a control chart for C, *i.e.*, number of defectives from the following data pertaining to the number of imperfections in 20 pieces of cloth of equal length in a certain make of polyester and infer whether the process is in a state of control :

2, 3, 5, 8, 12, 2, 3, 4, 6, 5, 6, 10, 4, 6, 5, 7, 4, 9, 7, 3.

Solution. Let C denote the number of defects per piece.

$$\bar{C} = \frac{\Sigma C}{N}$$

$$\Sigma C = 2 + 3 + 5 + 8 + 12 + 2 + 3 + 4 + 6 + 5 + 6 + 10 + 4 + 6 + 5 + 7 + 4 + 9 + 7 + 3 = 111$$

$$\bar{C} = \frac{111}{20} = 5.55$$

$$\text{UCL} = \bar{C} + 3\sqrt{\bar{C}} = 5.55 + 3\sqrt{5.55}$$
$$= 5.55 + 7.07 = 12.62.$$

$$\text{LCL} = \bar{C} - 3\sqrt{\bar{C}} = 5.55 - 7.07$$
$$= -1.52 \text{ or } 0$$

Since none of the points is falling outside the upper and lower control limits, the process is in a state of control.

Illustration 17. From a transistor production line 20 samples (each sample of 100 transistors) is chosen. The number of defects in each sample are given below :

Sample No.	*No. of defects*	*Sample No.*	*No. of defects*
1	44	11	36
2	48	12	52
3	32	13	35
4	50	14	41
5	29	15	42
6	31	16	30
7	46	17	46
8	52	18	38
9	44	19	26
10	48	20	30

Compute the values for an appropriate control chart and give your comments.

Solution. The appropriate control chart to be used is the C-chart. The computations required for preparing this chart are :

$\bar{C}$, *i.e.*, average no. of defects $= \frac{800}{20} = 40$

$$\text{UCL} = \bar{C} + 3\sqrt{\bar{C}} = 40 + 3\sqrt{40}$$
$$= 40 + (3 \times 6.324) = 40 + 18.972 = 58.972$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}} = 40 - 3\sqrt{40}$$
$$= 40 - (3 \times 6.324) = 40 - 18.972 = 21.028$$

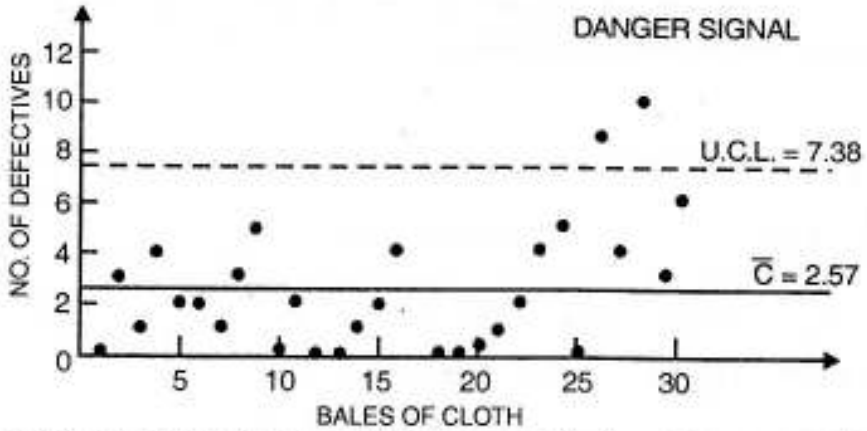
Since all the points are lying within control limits, the process is in a state of control.

Illustration 18. The following are the number of defects noted in the final inspection of 30 bales of woollen cloth : 0, 3, 1, 4, 2, 2, 1, 3, 5, 0, 2, 0, 0, 1, 2, 4, 3, 0, 0, 0, 1, 2, 4, 5, 0, 9, 4, 10, 3 and 6.
Compute the values for an appropriate control chart and give your comments. *(M.Com., Kurukshetra Univ., 1996)*

Solution. The appropriate chart would be a *C*-chart.
Total no. of defects = 87, sample size = 30

$$\bar{C} = \frac{77}{30} = 2.57$$
$$UCL = \bar{C} + 3\sqrt{\bar{C}} = 2.57 + 3\sqrt{2.57}$$
$$= 2.57 + 4.81 = 7.38$$
$$LCL = \bar{C} - 3\sqrt{\bar{C}} = 2.57 - 3\sqrt{2.57}$$
$$= 2.57 - 4.81 = -2.24 \text{ or } 0$$

Since lower control limit cannot be negative, we would start with zero.
The control chart is given below :



It is clear from the graph that two points are lying outside the control limits which represent danger signal.

Illustration 19. Samples of 100 tubes are drawn randomly from the output of a process that produces several thousand units daily. Sample tubes are inspected for quality and defective tubes are rejected. The results of 15 samples are shown below :

Sample No.	*No. of defective tubes*	*Sample No.*	*No. of defective tubes*
1	8	9	10
2	10	10	13
3	13	11	18
4	9	12	15
5	8	13	12
6	10	14	14
7	14	15	9
8	6		

On the basis of information given above prepare a control chart for fraction defective.

Solution. CONSTRUCTING CONTROL CHART FOR FRACTION DEFECTIVE

Sample No.	*No. of defectives*	*Fraction defectives*	*Sample No.*	*No. of defectives*	*Fraction defectives*
1	8	0.08	9	10	0.10
2	10	0.10	10	13	0.13
3	13	0.13	11	18	1.18
4	9	0.09	12	15	0.15
5	8	0.08	13	12	0.12
6	10	0.10	14	14	0.14
7	14	0.14	15	9	0.09
8	6	0.06			
			Total	169	

The suitable chart here will be the p-chart.

(*i*) Average fraction defective

$$\bar{p} = \frac{169}{1500} = 0.113$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= 0.113 + 3\sqrt{\frac{0.113(1-0.113)}{100}} = 0.113 + 0.095 = 0.208.$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= 0.113 - 3\sqrt{\frac{0.113(1-0.113)}{100}} = 0.113 - 0.095 = 0.018.$$

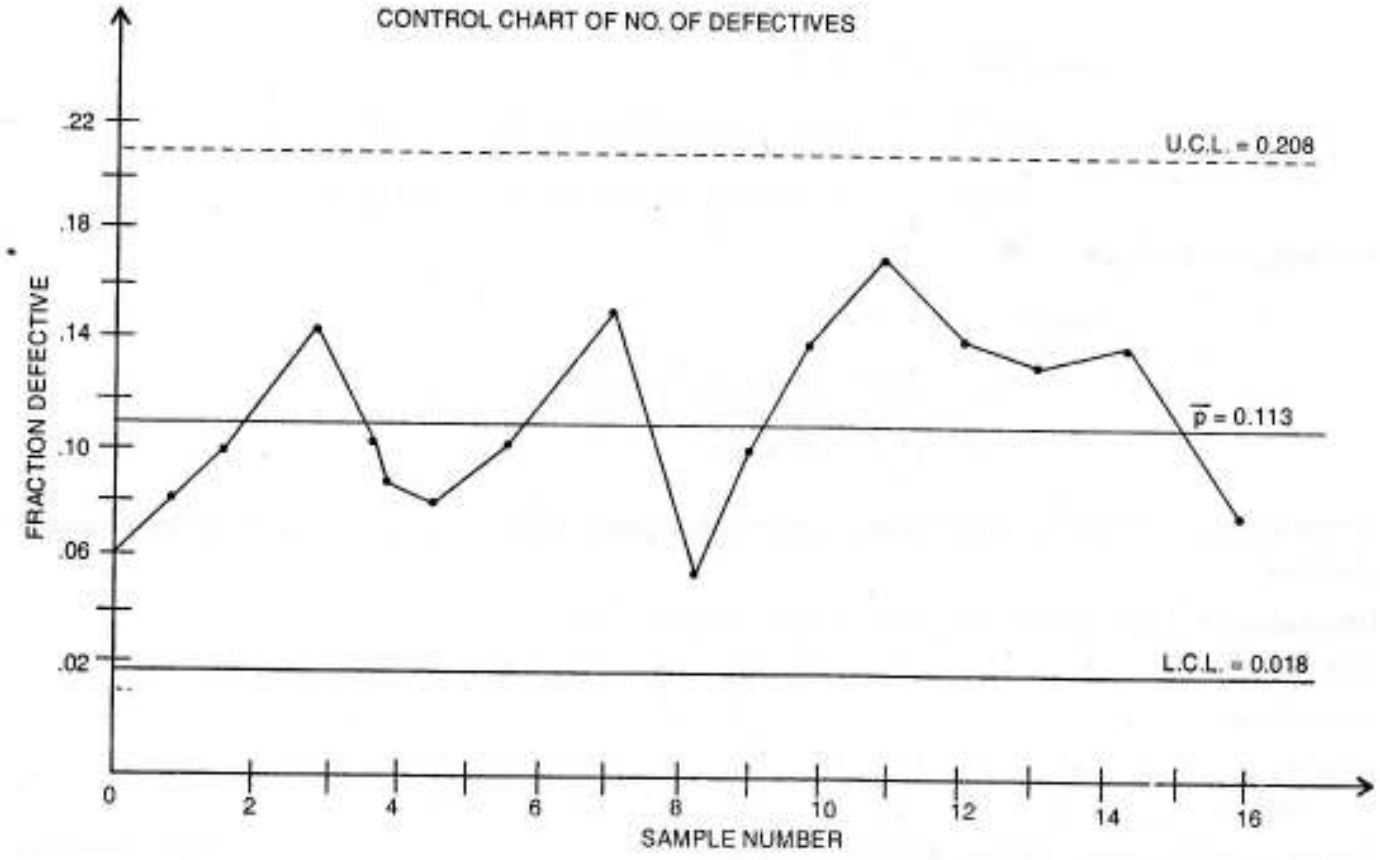


Illustration 20. Ten samples each of size 5 are drawn at regular intervals from a manufacturing process. The sample means ($\bar{X}$) and their average (R) are given below :

Sample No :	1	2	3	4	5	6	7	8	9	10
Mean ($\bar{X}$) :	49	45	48	53	39	47	46	39	51	45
Range (R) :	7	5	7	9	5	8	8	6	7	6

Calculate the control limits in respect of $\bar{X}$ -chart and R-chart.

(You are given : $A_2 = 0.58$, $D_3 = 0$, $D_4 = 2.115$). Comment on the state of control, charts need not be drawn.

Solution. (*a*) Control limits for $\overline{X}$ -chart.

Sample No.	*Mean*	*Range*
1	49	7
2	45	5
3	48	7
4	53	9
5	39	5
6	47	8
7	46	8
8	39	6
9	51	7
10	45	6
	$\Sigma \overline{X} = 462$	$\Sigma R = 68$

$$\overline{\overline{X}} = \frac{462}{10} = 46.2, \quad \overline{R} = \frac{68}{10} = 6.8$$

Central Line = $\overline{\overline{X}} = 46.2.$

$$\text{UCL} = \overline{\overline{X}} + A_2 \overline{R} = 46.2 + 0.58\,(6.8) = 46.2 + 3.944 = 50.144$$

$$\text{LCL} = \overline{\overline{X}} - A_2 \overline{R} = 46.2 - 0.58(6.8) = 46.2 - 3.944 = 42.256$$

Control limits for R-Chart

Central Line = $\overline{R} = 6.8$

$$\text{UCL} = D_4 \overline{R} = 2.115(6.8) = 14.382$$

$$\text{LCL} = D_3 \overline{R} = 0(6.8) = 0$$

In both cases, *i.e.*, $\overline{X}$ and $\overline{R}$ -charts, some of the points are lying outside the control limits. Hence the process is not in a state of control.

Illustration 21. The number of defects on 20 items are given below :

Item No.	:	1	2	3	4	5	6	7	8	9	10
No. of defects	:	2	0	4	1	0	8	0	1	2	0
Item No.	:	11	12	13	14	15	16	17	18	19	20
No. of defects	:	6	0	2	1	0	3	2	1	0	2

Prepare a suitable control chart and draw your conclusion. ***(MBA, Pune Univ., 2004)***

Solution. Let C denote the number of defects per item.

$$\overline{C} = \frac{\Sigma C}{n} = \frac{35}{20} = 1.75$$

$$\text{UCL} = \overline{C} + 3\sqrt{\overline{C}}$$

$$= 1.75 + 3\sqrt{1.75}$$

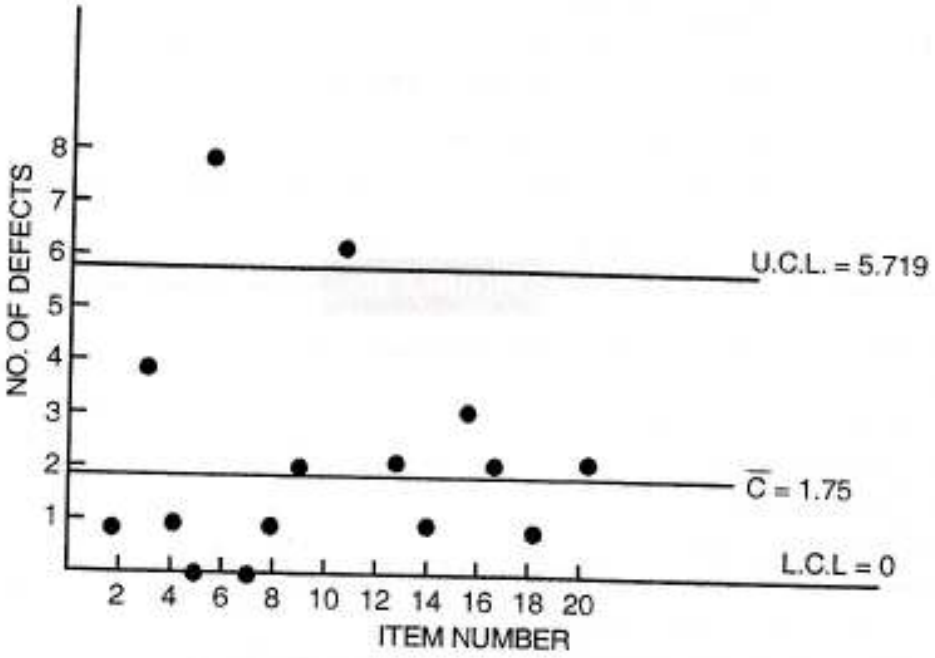
$$= 1.75 + 3 \times 1.323 = 5.719$$

$$\text{LCL} = \overline{C} - 3\sqrt{\overline{C}}$$

$$= 1.75 - 3\sqrt{1.75}$$

$$= 1.75 - 3 \times 1.323 = -2.219 \text{ or } 0$$

The control chart based on these limits is given below :



Since two of the points are lying outside the control limits, the process is not in a state of control.

Illustration 22. The following figures give the number of defectives in 20 samples containing 2000 items :

425, 430, 216, 341, 225, 322, 280, 306, 337, 305,
356, 402, 216, 264, 126, 409, 193, 280, 389, 326.

Find the control limits for the appropriate chart to be used.

Solution. The appropriate chart to be used in this case is a p-chart. Total number of defectives out of 40,000 items in 20 samples is :

$$= 425 + 430 + 216 + \ldots\ldots\ldots = 6148$$

$$\bar{p} = \frac{6148}{40000} = 0.1537$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}}$$

$$= 0.1537 - 3\sqrt{\frac{0.1537 \times 0.8463}{2000}} = 0.1537 - 0.0244 = 0.1293$$

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}} = 0.1537 + 0.0244 = 0.1781.$$

Illustration 23. You are given the values of sample means ($\bar{X}$) and the range (R) for the samples of size 5 each. Calculate the values for the mean and range control charts and comment on the state of control.

Sample No. :	1	2	3	4	5	6	7	8	9	10
$\bar{X}$:	43	49	37	4	5	37	51	46	43	47
R :	5	6	5	7	7	4	8	6	4	6

You may use the following control chart constants for $n = 5$, $A_2 = 0.58$, $D_3 = 0$, $D_4 = 2.115$. *(M.Com., DU, 2006)*

Solution. Control limits for $\bar{X}$ chart :

$$\text{CL} = \frac{\Sigma\bar{X}}{N} = \frac{442}{10} = 44.2; \quad \bar{\bar{X}} = \frac{58}{10} = 5.8$$

$$\text{UCL} = \bar{\bar{X}} + A_2\,\bar{R}$$
$$= 44.2 + 0.58(5.8) = 44.2 + 3.364 = 47.564$$

$$\text{LCL} = \bar{\bar{X}} - A_2\,\bar{R}$$
$$= 44.2 - 0.58(5.8) = 44.2 - 3.364 = 40.836$$

Since some of the points are lying outside the control limits, the process is not in a state of control.

Control limits for R chart :

$$\text{CL} = \bar{R} = \frac{58}{10} = 5.8$$

$$\text{UCL} = D_4 \bar{R} = 2.115(5.8) = 12.267$$

$$\text{LCL} = D_3 \bar{R} = 0(5.8) = 0$$

Since none of the points is lying outside the control limits, the range is in a state of control.

PROBLEMS

1-A: Answer the following questions, each question carries **one** mark:

(*i*) What is SQC ?
(*ii*) Give two important uses of SQC.
(*iii*) What is Control Chart ?
(*iv*) What is process control ?
(*v*) Why R-Chart is prepared ?
(*vi*) What are the control limits for p-chart ?
(*vii*) Is 100% inspection totally reliable ?
(*viii*) How do control charts reveal that the process is out of control ?
(*ix*) Distinguish between defects and defectives.
(*x*) What is OC curve ?
(*xi*) Define the terms 'AQL' and 'RQL'.

1-B: Answer the following questions, each question carries **four** marks:

(*i*) Differentiate between control chart of variables and attributes.
(*ii*) Explain the terms 'chance variation' and 'assignable variation' with suitable example.
(*iii*) What is $\bar{X}$ -chart ? How are the control limits determined while drawing this chart ?
(*iv*) What is C-chart ? Point out its uses.
(*v*) What is acceptance sampling ? Point out its role in business decision-making.
(*vi*) Distinguish between single sampling and double sampling plans.

2. What is a statistical quality control ? Point out its importance in the industrial world. Also explain the role of control charts.

3. (*a*) Distinguish between the process control and product control.
(*b*) Distinguish between the control limits and tolerance limits.

4. What is a control chart ? Describe how a control chart is constructed and interpreted.

5. Discuss the basic principles underlying control charts. Explain in brief, the construction and use of p-chart and C-chart.

6. What is control chart ? Explain in brief, the construction and use of mean chart, p-chart and range chart.

7. (*a*) What is acceptance sampling ? Point out the role of operating characteristic curve.
(*b*) Critically examine the different types of acceptance sampling plans.

8. (*a*) What do you mean by SQC? Discuss briefly its need and utility in industry. Discuss the causes of variation in quality.
(MBA, Vikram Univ., 2004)

(*b*) What are the various types of control charts known to you ? Explain them with examples.

9. "Quality control is attained most efficiently of course, not by the inspection operation itself but by getting at the causes." Comment on the statement. Describe the various devices employed for the maintenance of quality in a uniform flow of manufactured products.

10. Describe control charts for $\bar{X}$ and σ and derive expression for their control limits. What are the advantages of σ-chart over the R-chart ?

11. Explain the term "Statistical quality control". How is the process control achieved with the help of control chart ? What are the fundamentals underlying the construction of quality control chart ?

12. (*a*) Describe how a control chart for fraction defective is set ? What modification is needed if varying numbers are inspected on different occasions ?
(*b*) Discuss the role of C-chart in statistical quality control.

13. Explain the following terms occurring in sampling inspection plans :
(*a*) A.O.Q.L, (*b*) lot tolerance per cent defectives, (*c*) producer's risk, and (*d*) consumer's risk.

14. (*a*) Explain what are chance causes and assignable causes of variation in the quality of manufactured product.
(*b*) Assuming the characteristic variable follows a normal distribution (mean and standard deviation unknown), specify the control limits and the central line for the mean and range charts.

15. (*a*) Distinguish between process control and product control.
(*b*) State the different types of acceptance sampling plans.

16. (*a*) "The control charts make it possible to distinguish between those variations which are due to chance causes and those due to assignable causes".
Explain the terms 'chance causes' and 'assignable causes' and elucidate the statement.
(*b*) Distinguish between :
(*i*) Chance causes and assignable causes of variation.
(*ii*) Defect and defectives.
(*iii*) Control charts for variables and control charts for attributes.

17. What is the mathematical justification on which the control limits in $\overline{X}$-charts are set up? What is the purpose of an $\overline{X}$ control chart ?

18. Explain how a control chart helps to control the quality of a manufactured product. Describe the basis of control chart. Distinguish clearly between the charts for variables and charts for attributes.

19. (*a*) Why $\overline{X}$ and R-charts should be used simultaneously ? Justify with the help of an example.
(*b*) Explain *OC* curve. Also explain how various points on the curve are calculated, *i.e.*, show calculations for any point (not for $p = 0$).
(*c*) Discuss the uses of statistical quality control and control charts.

20. (*a*) What is an '*OC*" curve ? Which *OC* curve would be called ideal?
(*b*) Draw *OC* curve for the following single sampling plan.
$N = 50, n = 10, C = 1$.
(*c*) Write a short note on *C*-chart.

21. (*a*) Explain the construction and function of
(*i*) $\overline{X}$-chart (*ii*) R chart.
(*b*) State the advantages of quality control.
(*c*) Explain the construction of double sampling plan.
(*d*) Differentiate between p-chart and *C*-chart in context of statistical quality control.

22. (*a*) Distinguish between random variations and assignable variations. How is the distinction relevant in statistical quality control ?
(M.Com., DU, 1999)
(*b*) 25 sub-groups of 5 items each were taken in the measurement of an important dimension of a manufactured part. The mean of the 25 sub-groups was 0.6000 inches and the sum of the ranges of the sub-groups was 0.5 inches. Find the upper and lower control limits for the control chart for $\overline{X}$ and R.

23. The following data are the results of life tests on 15 samples of 6 fluorescent lamps each. The values are in hours.

Sample No.	$\overline{X}$	R	*Sample No.*	$\overline{X}$	R
1	4209	450	9	4420	320
2	4380	390	10	4385	510
3	4560	480	11	4182	490
4	3490	330	12	4260	385
5	3360	460	13	4550	220
6	3450	380	14	3890	490
7	3280	400	15	4280	160
8	3380	440			

(*a*) Is the process in a state of statistical quality control ?
(*b*) Assuming assignable causes could be discovered and eliminated, what is your best estimate of the capability of this process ?

24. A plant produces paper for newsprint, and rolls of paper are inspected for defects. The results of the inspection of 25 rolls of paper are given below :

Roll No.	*No. of defects*	*Roll No.*	*No. of defects*
1	10	14	5
2	20	15	4
3	8	16	2
4	12	17	3
5	13	18	6
6	15	19	8
7	25	20	9
8	7	21	15
9	13	22	18
10	18	23	20
11	16	24	10
12	14	25	5
13	6		

Draw control chart for defects and determine whether inspection results indicate stability.

25. Samples of 50 calculators are drawn randomly from the output of a process that produces several thousand units daily. Sampled items are inspected for quality, and faulty calculators are rejected. The result to a series of samples are given below :

Sample results of 15 lots of 50 calculators

Lot No.	*No. inspected*	*No. defectives*	*Lot No.*	*No. inspected*	*No. defectives*
1	50	4	9	50	5
2	50	5	10	50	6
3	50	8	11	50	8
4	50	10	12	50	5
5	50	6	13	50	12
6	50	7	14	50	4
7	50	3	15	50	2
8	50	2			

Draw a control chart and interpret it.

26. The number of defects found in inspecting television set assemblies are as follows for 20 inspection units of five sets each :

Unit	:	1	2	3	4	5	6	7	8	9	10
No. of defects	:	2	40	38	63	92	45	18	120	45	38
Unit	:	11	12	13	14	15	16	17	18	19	20
No. of defects	:	40	73	68	90	63	85	56	72	40	50

Set up a control chart to be used for future production.

27. A manufacturer purchases small bolts in cartons that usually contain several thousand bolts. Each shipment consists of a number of cartons. As part of the acceptance procedure for these bolts, 400 bolts are selected at random from each carton and are subjected to visual inspection for certain defects. In a shipment of 10 cartons, the respective percentages of defectives in the sample from each carton are 0.0, 0.5, 0.75, 0.20, 2.250, 0.25 and 1.25. Plot the appropriate control chart and draw your conclusions.

28. A machine is designed to produce ball bearings having a mean diameter of 0.574 cms and a standard deviation of 0.008 cms. To determine whether the machine is in proper working order, a sample of 6 ball bearings is taken every two hours on all the working days (namely Monday to Friday) of the week and the mean diameter is computed from this sample. Design a rule whereby one can be fairly certain that the quality of the products are conforming to required standards. Give a sketch of the control chart.

29. A sample of 200 bolts is drawn at regular intervals from the production line, and each bolt is checked. The number of defective bolts in 20 successive samples are given below :

Sample No.	*Defective bolts*	*Sample No.*	*Defective bolts*
1	3	11	2
2	3	12	3
3	1	13	2
4	3	14	1
5	2	15	1
6	3	16	3
7	2	17	3
8	2	18	3
9	3	19	2
10	3	20	3

Draw a suitable control chart and test whether the process is under control.

30. The following figures give the number of defectives in 20 samples ; each sample containing 2,000 items :

425, 430, 216, 341, 225, 322, 280, 306, 337, 305, 356, 402, 216, 264, 126, 409, 193, 326, 210, 389.

Calculate the values for central line and the control limits for *p*-chart (Fraction Defective chart). Draw the *p*-chart and comment if the process can be regarded in control or not.

31. With a view to examine the quality of an engineering product, 10 samples of 200 items each were taken from a day's production and the number of defective items in each sample was recorded as follows :

Sample No.	1	2	3	4	5	6	7	8	9	10
No. of defectives	14	20	36	42	22	18	26	2	12	8

(*i*) Draw a fraction defective quality control chart, showing clearly the upper and lower control limits.

(*ii*) What inference do you draw from this quality control chart ?

32. A large sample of a product gave an "average fraction defective" of 0.068. Calculate for a *p*-chart the values of control limits (upper, lower and central line), if the size of each sample sub-group is 200.

33. In a glass factory, the task of quality control was done with the help of mean ($\overline{X}$) and standard deviation (σ) charts. 18 samples of 10 items each were chosen and their values ΣX and $\Sigma\sigma$ were found to be 595.1 and 8.28 respectively. Determine 3σ limits for mean and standard deviation charts. You may use the following control factors for your calculations :

n	A_1	B_2	B_3
10	1.03	0.28	1.72

34. In a factory that produces steel tubes, the thickness of walls is to be controlled. Every hour a sample of 6 tubes is taken and after measurements average thickness in centimetres and the range for each sample is noted.

Sample No. :	1	2	3	4	5	6	7	8	9	10
Average thickness :	0.25	0.32	0.42	0.22	0.28	0.10	0.25	0.40	0.06	0.29
Range :	0.25	0.48	0.12	0.12	0.19	0.10	0.06	0.46	0.10	0.32

Draw average and range charts and give your comments whether the process is under control or not.

35. (*a*) Draw an *OC* curve for the following sampling plan, which is used to inspect lots of size 500 items each.

Sample size = 10; Acceptance No. = 1; Rejection No. = 2

(*b*) Describe briefly a multiple acceptance sampling plan.

36. Control on measurements of pitch diameter of thread in aircraft fittings is checked with 5 successive items measured at regular intervals, 5 such samples are given below :

Sample	*Measurement on each item of 5 items per hour*				
1	45	45	44	43	42
2	41	41	44	42	40
3	40	40	42	40	42
4	42	43	42	42	45
5	43	44	47	47	45

(Values are expressed in units of 0.001 inch.)

(*i*) Construct the $\overline{X}$ and *R*-charts.

(*ii*) What inference do you draw from these quality control charts ?

($n = 5, A_1 = 0546, A_2 = 0.577, D_2 = 4.981, D_3 = 0, D_4 = 2.115$)

37. It has been ascertained that when a manufacturing process is under control, the average of the defectives per sample batch of 10 is 12. What limits would you set in a quality control chart based on the examination of defectives in sample batches of 10 ?

38. Process for producing solid state devices such as transistors frequently have a rather high fraction defective. One particular line for making transistors has a long run fraction defective of 0.28 when functioning properly. Every two hours a sample of 50 transistors is examined and the number of defectives in the sample determined. What are the control limits for the *p*-chart used to control this process ?

39. The Quality Electric Bulbs Ltd. manufactures electric bulbs under an improved process. The Production Engineer takes a random sample of 100 bulbs off the run from each day's output for inspection. The number of defective pieces is determined by applying a high voltage test. Suppose that the process of manufacture when under control admits of a long run fraction defectives of 0.05 ?Determine the control limits on the *p*-chart.

40. When will you use a control chart for defects ? Plot control chart for the following data pertaining to number of defects in the calculators manufactured by a company :

Calculator No. :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of Defects :	5	1	0	7	3	6	0	10	2	11	5	8	6	4	1

41. Given below is a record of the number of defects seen in circuit panels used in a computer.

Prepare an appropriate control chart to control the quality of the product. Assess whether the process is in control as per data observed.

If assignable causes have been found to explain outliers, how does this affect the control chart. Show the necessary changes in control limits.

Panel	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
No. of defects	:	4	0	1	4	5	3	6	2	2	0	5	10	3	4

42. The following data refers to visual defect found during the inspection of the first 10 samples of size 50 each from a lot of Two-wheelers manufactured by an Automobile Company :

Sample No.	:	1	2	3	4	5	6	7	8	9	10
No. of Defectives	:	4	3	2	3	4	4	4	1	3	2

Draw the '*p*' chart to show that the fraction defectives are under control.

[$UCL = 0.1608; LCL = 0$]

43. The following data show the values of sample mean and the range for ten samples of size 5 each. Construct the $\overline{X}$ and Range charts :

Sample No. :	1	2	3	4	5	6	7	8	9	10
Mean ($\overline{X}$) :	11.2	11.8	10.8	11.6	11.0	9.6	10.4	9.6	10.6	10.0
Range (R) :	7	4	8	5	7	4	8	4	7	9

44. The average number of defectives, in 22 sampled lots of 2000 rubber belts each, was found to be 16%. Determine the 3-sigma control limits for the *p*-chart. ***(Diploma in Mgt., AIMA, 2004)***

45. The following data shows the mean and the range for ten samples of size each. Calculate the values for the central line and control limits for mean-chart and range-chart, and determine whether the process is in control.

Sample No. :	1	2	3	4	5	6	7	8	9	10
Mean :	11.4	12.0	11.0	11.8	11.2	9.8	10.6	9.8	10.8	10.2
Range :	7	4	8	5	7	4	8	4	7	9

(M.Com., DU, 2006)

*****

20

Partial and Multiple Correlation and Regression

INTRODUCTION

The simple correlation and regression analysis discussed earlier measure the degree and nature of the effect of one variable on another. While it is useful to know how one phenomenon is influenced by another, it is also important to know how one phenomenon is affected by several other variables. One variable is related to a number of other variables, many of which may be interrelated among themselves. For example, yield of rice is affected by the type of soil, temperature, amount of rainfall, etc. It is part of the statistician's task to determine the effect of one case when the effect of others is estimated. This is done with the help of multiple and partial correlation analysis. Thus, it shall be possible for us to compare the relative importance of television advertisement and newspaper advertisement on increasing sales.

The basic distinction between multiple and partial correlation analysis is that whereas in the former, we measure the degree of association between the variable Y and all the variables, $X_1, X_2, X_3, \ldots X_n$, *taken together*; in the latter we measure the degree of association between Y and one of the variables $X_1, X_2, X_3, \ldots X_n$, *with the effect of all the other variables removed.* It should be noted that when only two variables are included in a study, the dependent variable is usually designated by Y, and the independent variable by X. However, when more than one independent variable is used it becomes advantageous to distinguish between the variables by means of subscripts and use only the letter X. The dependent variable is generally denoted by X_1 and the independent variables by X_2, X_3, etc. This scheme of notation can be expanded to take care of any number of independent variables.

PARTIAL CORRELATION

It is often important to measure the correlation between a dependent variable and one particular independent variable when all other variables involved are kept constant, *i.e.,* when the effects of all other variables are removed (often indicated by the phrase "other things being equal"). This can be obtained by calculating coefficient of partial correlation. For example, if we have three variables : yield of wheat, amount of rainfall and temperature and if we limit our analysis of yield and rainfall to periods when a certain average daily temperature existed, or if we treat the problem mathematically in such a way that changes in temperature are allowed for, the problem becomes one of partial correlation. Thus, partial correlation analysis measures the strength of the relationship between Y and one independent variable in such a way that variations in the other independent variables are taken into account. A partial correlation coefficient is analogous to a partial regression coefficient in that all other factors are "*held constant*". Simple correlation, on the other hand, ignores the effect of all the other variables even through these variables might be quite closely related to the dependent variable, or to one another.

Partial Correlation Coefficients

Partial correlation coefficients provide a measure of the relationship between the dependent variable and other variables, with the effect of the rest of the variables eliminated.

If we denote by $r_{12.3}$, the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant, we find that

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

where, $r_{13.2}$ is the coefficient of partial correlation between X_1 and X_3 keeping X_2 constant.

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}$$

where $r_{23.1}$ is the coefficient of partial correlation between X_2 and X_3 keeping X_1 constant. Thus, for three variables X_1, X_2 and X_3, there will be three coefficients of partial correlation each studying the relationship between two variables when the third is held constant. It should be noted that the squares of partial correlation coefficients are called *coefficients of partial determination.*

It will be clear from above that the partial correlation coefficients measure the degree of correlation between the dependent variable and each independent variable when the values of specified combinations of the other independent variables are held constant. These coefficients enable us to determine the direct relationship between any two variables independent of the indirect effect of the other variables : Partial correlation coefficient helps in deciding whether to include or not, an additional independent variable in regression analysis. Depending on the number of independent variables held constant, we often talk of *zero-order, first-order, second-order* correlation coefficients. The correlation coefficients obtained in case of two variables, *i.e.,* X and Y is called *zero-order correlation coefficient* since no restrictions are imposed on the values of all variables other than X and Y. The zero-order coefficients possess no secondary subscripts—that is subscripts after the point. If one independent variable is held constant in correlating two other variables, the resulting coefficient is known as the *first-order correlation* coefficient. Thus in a trivariate case, the partial correlation and regression coefficients such as $r_{12.3}$, $r_{13.2}$, $r_{23.1}$ are called first order coefficients. In a similar manner, a correlation between two variables while holding the values of the two other variables constant is known as a *second-order correlation coefficient.* Examples of second order coefficient are $r_{12.23}$, $r_{14.23}$, $r_{13.24}$, etc.

The notation of partial correlation coefficients always follows the same principle ; namely, the two variables being correlated are identified by the subscripts of r before the dot and the variables held constant are identified by the subscripts after the dot. It may be noted that so long as the particular subscripts of r are on the correct side of the period, the order in which they are placed is of no consequence. For example, $r_{12.34} = r_{12.43} = r_{21.43}$. However, the usual practice among statisticians is to place the subscripts in the ascending order.

Coefficients of a given order can generally be expressed in terms of the next lower order such as expressing partial correlations for the trivariate case in terms of simple correlations. This possibility simplifies greatly, the computational work involved in case of three or four independent variables.

A partial correlation coefficient measures the net co-variation between two of the variables under consideration. It is interpreted in terms of its squared values—coefficient of partial determination. For

instance, if $r_{12.3} = 0.8$ then $r^2_{12.3}$ would be 0.64 which means that the errors made in estimating X_1 from X_2 are reduced by 64 per cent when X_3 is employed as an additional explanatory variable.

Illustration 1. In a trivariate distribution it is found that

$$r_{12} = 0.7,\ r_{13} = 0.61, r_{23} = 0.4$$

Find the values of $r_{23.1}$, $r_{13.2}$ and $r_{12.3}$

Solution.
$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{13}^2}}$$

Substituting the given values

$$r_{23.1} = \frac{0.4 - 0.7 \times 0.61}{\sqrt{1-(0.7)^2}\ \sqrt{1-(0.61)^2}} = \frac{0.4 - 0.427}{\sqrt{0.51}\ \sqrt{1-0.3721}} = 0.048$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{23}^2}}$$

$$= \frac{0.61 - 0.7 \times 0.4}{\sqrt{1-(0.7)^2}\ \sqrt{1-(0.4)^2}} = \frac{0.61 - 0.28}{\sqrt{1-0.49}\ \sqrt{1-0.16}} = 0.504$$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}} = \frac{0.7 - 0.6 \times 0.4}{\sqrt{1-(0.61)^2}\ \sqrt{1-(0.4)^2}} = 0.633$$

Illustration 2. On the basis of observation made on agricultural production (X_1) the use of fertilizers (X_2) and the use of irrigation (X_3), the following zero order correlation coefficients were obtained :

$$r_{12} = 0.8, r_{13} = 0.65, r_{23} = 0.7$$

Compute the partial correlation between agricultural production and the use of fertilizers eliminating the effect of irrigation.

Solution. We have to calculate the value of $r_{12.3}$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}}$$

Substituting the values $r_{12.3} = 0.8, r_{13} = 0.65$ and $r_{23} = 0.7$

$$r_{12.3} = \frac{0.8 - (0.65 \times 0.7)}{\sqrt{1-(0.65)^2}\ \sqrt{1-(0.7)^2}} = \frac{0.8 - 0.455}{\sqrt{1-0.4225}\ \sqrt{1-0.49}} = 0.636.$$

Illustration 3. Is it possible to get the following from a set of experimental data.

(*a*) $r_{23} = 0.8, r_{31} = -0.5, r_{12} = 0.6$ (*b*) $r_{23} = 0.7,\ r_{31} = -0.4, r_{12} = 0.6.$

Solution. In order to see whether there is any inconsistency, we should calculate $r_{12.3}$. If its value exceeds one, there is inconsistency otherwise not.

(*a*)
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}} = \frac{0.6 - (-0.5)\,(0.8)}{\sqrt{1-(-0.5)^2}\ \sqrt{1-(0.8)^2}}$$

$$= \frac{0.6 + 0.4}{\sqrt{0.75}\ \sqrt{0.36}} = \frac{1}{0.52} = 1.92$$

Since the value of $r_{12.3}$ is greater than one, therefore, there is some inconsistency in the given data.

(*b*)
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}}$$

$$= \frac{0.6 - (-0.4)\,(0.7)}{\sqrt{1-(0.4)^2}\ \sqrt{1-(0.7)^2}} = \frac{0.6 + 0.28}{\sqrt{0.84}\ \sqrt{0.51}} = \frac{0.88}{0.65} = 1.35$$

This again is greater than one, therefore, there is some inconsistency in the given data.

Partial Correlation Coefficients in more than three variables

When four variables are involved in a correlation problem, there are twelve possible first-order coefficients. Some of these are :

$$r_{14.2} = \frac{r_{14} - r_{12}\, r_{24}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{14.3} = \frac{r_{14} - r_{13}\, r_{34}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{13.4} = \frac{r_{13} - r_{14}\, r_{34}}{\sqrt{1 - r_{14}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{12.4} = \frac{r_{12} - r_{14}\, r_{24}}{\sqrt{1 - r_{14}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{24.3} = \frac{r_{24} - r_{23}\, r_{34}}{\sqrt{1 - r_{23}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{34.2} = \frac{r_{34} - r_{23}\, r_{24}}{\sqrt{1 - r_{23}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{23.4} = \frac{r_{23} - r_{24}\, r_{34}}{\sqrt{1 - r_{24}^2}\ \sqrt{1 - r_{34}^2}}.$$

Similarly, the formulae for other partial correlation coefficients, *i.e.*, $r_{12.3}$, $r_{13.2}$, $r_{23.1}$, $r_{24.1}$, $r_{34.1}$, can also be written.

Second-order Partial Correlation Coefficients

Second-order coefficients may be obtained from order coefficients. In case of four variables, if $r_{12.34}$ is the coefficient of partial correlation between X_1 and X_2 keeping X_3 and X_4 constant, then

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}\, r_{23.4}}{\sqrt{1 - r_{13.4}^2}\ \sqrt{1 - r_{23.4}^2}}.$$

Similarly,

$$r_{13.24} = \frac{r_{13.4} - r_{12.4}\, r_{23.4}}{\sqrt{1 - r_{12.4}^2}\ \sqrt{1 - r_{23.4}^2}}$$

and

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}\, r_{24.3}}{\sqrt{1 - r_{12.3}^2}\ \sqrt{1 - r_{24.3}^2}}.$$

Alternative formulae giving the same results are available for all three of the second-order coefficients. They are :

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{1 - r_{14.3}^2}\ \sqrt{1 - r_{24.3}^2}}$$

$$r_{13.24} = \frac{r_{13.2} - r_{14.2}\, r_{34.2}}{\sqrt{1 - r_{14.2}^2}\ \sqrt{1 - r_{34.2}^2}}$$

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}\, r_{34.2}}{\sqrt{1 - r_{13.2}^2}\ \sqrt{1 - r_{34.2}^2}}.$$

The value of a partial correlation coefficient is usually interpreted via the corresponding coefficient of partial determination, which is merely the square of the former. Thus, if $r_{12.3} = 0.4$, $r^2_{12.3} = 0.16$.

The *t*-test employed to test the significance of a simple correlation can be employed to test the significance of a partial correlation when the number of degrees of freedom is reduced by the number of variables eliminated.

Characteristics and Uses of Partial Correlation Analysis. The purpose of partial correlation analysis is the measurement of relationship between two factors, with the effects of one or more other factors eliminated. If the assumptions of the method are true for a series of data, the power of partial analysis is great. The problem of holding certain variables constants while the relationship between the other is measured often presents itself in statistical analysis. Partial correlation is especially useful in the analysis of interrelated series. It is particularly pertinent to uncontrolled experiments of various kinds in which, such interrelationship usually exists. Most economic data fall in this category.

Partial correlation is of greatest value when used in conjunction with gross and multiple correlation in the analysis of factors affecting variations in many kinds of phenomena.

It has the advantage that the relationships are expressed concisely in a few well-defined coefficients. Also it is adaptable to small amounts of data and the reliability of the results can be rather easily tested.

Limitations of Partial Correlation Analysis. 1. The usefulness of the partial analysis is somewhat limited by the following basic assumptions of the method :

(*i*) The gross or zero-order correlation must have linear regressions.

(*ii*) The effects of the independent variable must be additively and not jointly related.

(*iii*) Because the reliability of partial coefficients decreases as its order increases. The number of observations in gross correlations should be fairly large. Often the student carries the analysis beyond the limits of the data. Thus, weakness to some extent can be guarded against by test of reliability.

2. When the above assumptions have been satisfied, partial correlation analysis still possess the disadvantages of laborious calculations and difficult interpretation even for statisticians.

The interpretation of the partial and multiple correlation results tends to assume that the independent variable have casual effects on dependent variable. The assumption is sometimes true, but more often untrue in varying degrees.

MULTIPLE CORRELATION

In problems of multiple correlation, we are dealing with situations that involve three or more variables. For example, we may consider the association between the yield of wheat per acre and both the amount of rainfall and the average daily temperature. We are trying to make estimates of the value of one of these variables based on the values of all the others. The variable whose value we are trying to estimate is called the dependent variable and the other variables on which our estimates are based are known as independent variables. The statistician himself chooses which variable is to be dependent and which variables are to be independent. It is merely a question of problem being studied. If we are trying to determine the most probable weight of men, we make weight, the dependent variable and height, age, etc., independent variables. If on the other hand, we are interested in estimating height, we will make height the dependent variable and weight, age, etc., the independent variables. Thus in problems of multiple correlation, we always have three or more variables (one dependent and the rest independent). In order that we may distinguish them easily, we follow the custom of representing them by the letter X with subscript. The dependent variable is always denoted by X_1 and the others by X_2, X_3, etc. Thus in the height, age and weight problem, if we are trying to estimate men's weight (that is, if weight be dependent variable), we might denote

$X_1 \rightarrow$ weight in lbs.
$X_2 \rightarrow$ height in inches.
$X_3 \rightarrow$ age in years.

The multiple correlation is of great practical significance—for rarely is it ever true that a variable is influenced solely or predominantly by one other factor. For example, the sales of a manufacturer are influenced, among other things, by his prices, his competitive position in the industry, his sales promotion campaign, industry sales, competitors' prices and national prosperity. In simple correlation, only one of the independent variables at a time could be correlated with the manufact ırer's sales and there is no direct way of determining the extent to which the observed correlation might have been caused by the interacting influence of other factors on the two variables under study. For instance, in times of prosperity a high level of national income may lead to increased industry sales, a share of which is captured by this manufacturer. But to what extent are the manufacturer's sales influenced by the universally buoyant affect of national prosperity and to what extent are his sales affected by the particular trend of the industry sales within the economy, *i.e.*, assuming that the nation's economy remain fairly stable ? To answer questions of this type which are of vital importance in framing suitable managerial policies, one has to depend on multiple correlation analysis.

Coefficient of Multiple Correlation

The coefficient of multiple linear correlation* is represented by R_1 and it is common to add subscript designating the variables involved. Thus $R_{1.234}$ would represent the coefficient of multiple linear correlation between X_1 on the one hand, and X_2, X_3 and X_4 on the other. The subscript of the dependent variable is always to the left of the point.

The coefficient of multiple correlation can be expressed in terms of r_{12}, r_{13} and r_{23} as follows :

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

$$R_{2.13} = \sqrt{\frac{r^2_{21} + r^2_{23} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{13}}}$$

and

$$R_{3.12} = \sqrt{\frac{r^2_{13} + r^2_{23} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{12}}}$$

A coefficient of multiple correlation such as $R_{1.23}$ lies between 0 and 1. The closer it is to 1, the better is the linear relationship between the variables. The closer it is to 0, the worse is the linear relationship. If the coefficient of multiple correlation is 1, the correlation is called *perfect*. Although a correlation coefficient of 0 indicates no linear relationship between the variables, it is possible that a nonlinear relationship may exist. It should be noted that whereas the simple correlation coefficients range from + 1.0 to 0 to – 1.0, the coefficients of multiple correlation are always positive in sign and range from + 1.0 to 0.

An alternative formula for calculating $R_{1.23}$ is :

$$R_{1.23} = \sqrt{r^2_{12} + r_{13.2}\,(1 - r^2_{12})}.$$

*When a linear regression equation is used, the coefficient of multiple correlation is called the coefficient of linear multiple correlation unless otherwise specified. Whenever we refer to multiple correlation, we shall imply linear multiple correlation.

similarly, $$R_{1.24} = \sqrt{\frac{r^2_{12} + r^2_{14} - 2r_{12}\, r_{14}\, r_{24}}{1 - r^2_{24}}}$$

or $$R_{1.24} = \sqrt{r^2_{12} + r^2_{14.2}\,(1 - r^2_{12})}$$

and $$R_{1.34} = \sqrt{\frac{r^2_{13} + r^2_{14} - 2r_{13}\, r_{14}\, r_{34}}{1 - r^2_{34}}}$$

or $$R_{1.34} = \sqrt{r^2_{13} + r^2_{14.3}\,(1 - r^2_{13})}.$$

Coefficient of Multiple Determination

In chapter on correlation, we talked of coefficient of determination r^2 which measures the fit of a straight line to the two-variable scatter. In exactly the same way, the coefficient of multiple determination denoted by $R^2_{1.23}$ is also defined. Thus, $R^2_{1.23}$ may be thought of as a measure of closeness of fit of the regression plane to the actual points relative to the point of the means of the variable. Or, just as does r^2, $R^2_{1.23}$ measures the percentage of total error that is accounted for by the regression. Obviously, the greater the value of $R^2_{1.23}$, the smaller is the scatter and the better is the fit. Thus, if coefficient of multiple determination between yield of rice (X_1) and fertilizers (X_2) and rain (X_3) is 0.953, it means that 95.3 per cent of the variations in yield have been explained by the variation in fertilizers and rain. There remains only 4.7 per cent of the variations in yield of rice that can be explained only by factors which have not been taken into consideration in our analysis.

Illustration. 4. The following zero-order, correlation coefficients are given

$$r_{12} = 0.98,\ r_{13} = 0.44 \text{ and } r_{23} = 0.54.$$

Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

Solution. We have to calculate the multiple correlation coefficient treating first variable as dependent and second and third variables as independent, *i.e.*, we have to find $R_{1.23}$

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

Substituting the given values

$$R_{1.23} = \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)\,(0.44)\,(0.54)}{1 - (0.54)^2}}$$

$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4657}{0.7084}} = +0.986.$$

Advantages of Multiple Correlation Analysis. The coefficient of multiple correlation serves the following purposes :

1. It serves as a measure of the degree of association between one variable taken as the dependent variable and a group of other variables taken as the independent variables.

2. Hence it also serves as a measure of goodness of fit of the calculated plane of regression and consequently, as a measure of the general degree of accuracy of estimates made by reference to equation for the plane or regression.

Limitations of Multiple Correlation Analysis. 1. Multiple correlation analysis is based on the assumption that the relationship between the variables is linear. In other words, the rate of change in one variable in terms of another is assumed to be constant for all values. In practice, most relationship are not linear but follow some other pattern. This limits somewhat the use of multiple correlation analysis. The linear regression coefficients are not accurately descriptive of curvilinear data.

2. A second important limitation is the assumption that effects of independent variables on the dependent variables are separate, distinct and additive. When the effects of variables are additive, a given change in one has the same effect on the dependent variable regardless of the sizes of the other two independent variables.

3. Linear multiple correlation involves a great deal of work relative to the results frequently obtained. When the results are obtained, only a few students well trained in the method are able to interest them. The misuse of correlation results has probably cast more doubt on the method than is justified. However, this lack of understanding and resulting misuses are due to the complexity of the method.

Multiple Regression

In the simple linear regression model discussed earlier, we talked of one dependent variable and one independent variable.

In multiple regression analysis which is a logical extension of two-variable regression analysis, instead of a single independent variable, two or more independent variables are used to estimate the values of a dependent variable. However, the fundamental concepts in the analysis remain the same. The multiple regression and correlation analysis serves highly useful purpose in practice. Its main objectives are :

(*a*) To derive an equation which provides estimates of the dependent variable from values of the two or more independent variables.

(*b*) To obtain a measure of the error involved in using this regression equation as a basis for estimation.

(*c*) To obtain a measure of the proportion of variance in the dependent variable accounted for or "explained by" the independent variables.

The first purpose is accomplished by deriving an appropriate regression equation by the method of least squares. The second purpose is achieved through the calculation of standard error of estimate and the third purpose is accomplished by computing the multiple coefficient of determination.

The multiple regression equation involving two independent variables shall take the form

$$Y_c = a + b_1X_1 + b_2X_2$$

The general form of the linear multiple regression function for k independent variables

$$X_1, X_2, \ldots\ldots, X_k, \text{ is}$$

$$Y_c = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_kX_k$$

The linear function which is fitted to data for two variables is referred to as a *straight line,* for three variables a *plane,* for four or more variables a *hyperplane.*

If we have three variables X_1, X_2 and X_3, the multiple linear regression equation of X_1 on X_2 and X_3 shall have the following form :

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Obviously X_1 is the dependent variable here and X_2 and X_3 are independent variables. The constant $a_{1.23}$ is the intercept made by the regression plane; it is zero when the regression line passes through the origin. The regression coefficients denoted by $b_{12.3}$ and $b_{13.2}$ represent the rate of change of the dependent variable per unit change in each of the independent variables when the other independent variables are held constant. The first subscript always represents the dependent variable and the second subscript denotes the particular independent variable being related to X_1. The subscripts after the period indicate the other independent variables, all of which are held constant while the effect of the particular independent variable on X_1 is measured. Thus, $b_{12.3}$ measures the amount by which a unit change in X_2 is expected to

affect X_1 when X_3 is held constant and $b_{12.3}$ measures the amount of change in X_1 when X_3 is held constant and $b_{13.2}$ measures the amount of change in X_1 per unit change in X_3 when X_2 is held constant. Similarly, $b_{13.24}$ would represent the change in X_1 per unit change in X_3 when the values of X_2 and X_4 are held constant.

The regression coefficients, *i.e.*, b's in multiple linear regression are termed *coefficients of net regression* : the regression is *net* in the sense that the regression of the dependent variable on the particular independent variable is measured while holding the values of the other independent variables constant. It contrast, the coefficients in simple regression are called *coefficients of gross regression* because no allowance is made for indirect influences on the regression.

The follwing are the usual assumptions made in a linear multiple regression analysis illustrated for the case of two independent variables :

1. The conditional distributions of Y for given X_1 and X_2 are assumed to be normal.
2. These conditional distributions are assumed to have equal standard deviations.
3. The $Y - Y_c$ deviation are assumed to be independent of one another.

Normal Equations for the Least Square Regression Plane

Just as there exist, least square regression line approximating a set of N data points (X, Y) in a two-dimensional scatter diagram, so also there exist *least square regression* planes fitting a set of N data points (X_1, X_2, X_3) in a three-dimensional by scatter diagram.

The least square regression plane of X_1 on X_2 and X_3 has the equation (i) where $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously, the *normal equations*.

$$\Sigma X_1 = Na_{1.23} + b_{12.3}\Sigma X_2 + b_{13.2}\Sigma X_3$$
$$\Sigma X_1 X_2 = a_{1.23}\Sigma X_2 + b_{12.3}\Sigma X_2^2 + b_{13.2} X_2 X_3$$
$$\Sigma X_1 X_3 = a_{1.23}\Sigma X_3 + b_{12.3}\Sigma X_2 X_3 + b_{13.2} X_3^2.$$

These can be obtained formally by multiplying both sides of equation (i) by 1, X_2 and X_3 successively and summing on both sides.

When the number of variables is 4 or more, solving the above system of normal equation becomes a very tedious procedure. Efficient methods solving simultaneous equations require a knowledge of matrix algebra, which is not assumed for the reader of the text. Thus in our discussion that follows, we shall confine ourselves to the two independent variables cases, which of course, can be extended to cover cases with three or more independent variables.

The work involved in finding these regression equations can be reduced by proceeding in terms of deviations from the mean of the variables under consideration. The regression equation for these variables in this procedure is :

$$x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3$$

where $x_1 = (X_1 - \bar{X}_1),\ x_2 = (X_2 - \bar{X}_2),\ x_3 = (X_3 - \bar{X}_3)$.

The value $b_{12.3}$ and $b_{13.2}$ can be obtained by solving simultaneously the following two normal equations :

$$\Sigma x_1 x_2 = b_{12.3}\Sigma x_2^2 + b_{13.2}\Sigma x_2 x_3$$
$$\Sigma x_1 x_3 = b_{12.3}\Sigma x_2 x_3 + b_{13.2}\Sigma x_3^2.$$

The value of $b_{12.3}$ and $b_{13.2}$ can also be obtained as follows :

$$b_{12.3} = r_{12.3}\,\frac{\sigma_{1.23}}{\sigma_{2.13}}$$

$$b_{13.2} = r_{13.2}\frac{\sigma_{13.2}}{\sigma_{3.12}}$$

The regression equation of X_1 on X_2 and X_3 can be expressed as follows :

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_3}\right)(X_3 - \bar{X}_3) \quad ...(i)$$

The regression equation of X_3 on X_2 and X_1 can be written as follows :

$$X_1 - \bar{X}_2 = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \bar{X}_1) \quad ...(ii)$$

From (*i*) and (*ii*), the coefficients of X_3 and X_1 are respectively

$$b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_2}\right) \text{ and } \left(\frac{r_{13} - r_{23}r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_2}{S_1}\right)$$

$$b_{13.2}\, b_{31.2} = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r^2_{23})(1 - r^2_{12})} = r^2_{13.2}$$

This method of obtaining regression equations is much simpler compared to one where simultaneously several normal equations are to be solved. For calculating regression equation for three variables, when the above procedure is used we need the following :

X_1	X_2	X_3
S_1	S_2	S_3
r_{12}	r_{13}	r_{23}

Other Equations of Multiple linear Regression

In the case of two variables, there were two equations of regression—one of them indicating regression of Y on X, and the other, that of X on Y. When there are three variables, there will be three equations of regression indicating the regression of X_1 on X_2 and X_3, the other indicating the regression of X_2 on X_1 and X_3 and the third indicating the regression of X_3 on X_1 and X_2. The first of these has been given earlier. If X_2 and X_3 were to be treated as dependent variables the regression equation will respectively be :

$$X_2 = a_{2.13} + b_{21.3}X_1 + b_{23.1}X_3 \quad ...(ii)$$

$$X_3 = a_{3.12} + b_{31.2}X_1 + b_{32.1}X_2 \quad ...(iii)$$

The normal equations for fitting (*ii*) will be :

$$\Sigma X_2 = Na_{2.13} + b_{21.3}\Sigma X_1 + b_{23.1}\Sigma X_3$$

$$\Sigma X_1X_2 = a_{2.13}\Sigma X_1 + b_{21.3}\Sigma X_1^2 + b_{23.1}\Sigma X_1X_2$$

$$\Sigma X_2X_3 = a_{2.13}\Sigma X_3 + b_{21.3}\Sigma X_1X_3 + b_{23.1}\Sigma X_3^2$$

In case we want to fit equation (*iii*), the normal equations will be :

$$\Sigma X_3 = Na_{3.12} + b_{31.2}\Sigma X_1 + b_{32.1}\Sigma X_2$$

$$\Sigma X_1X_3 = a_{3.12}\Sigma X_1 + b_{31.2}\Sigma X_1^2 + b_{32.1}\Sigma X_1X_2$$

$$\Sigma X_2X_3 = a_{3.12}\Sigma X_2 + b_{31.2}\Sigma X_1X_2 + b_{32.1}\Sigma X_2^2$$

Generalization for More Than Three Variables

In case of four variables, the linear regression equation of X_1 on X_2, X_3 and X_4 can be written as

$$X_1 = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

It represents a hyperplane in four-dimensional space. On formal multiplication of both sides of the above equation by X_1, X_2, X_3 and X_4 successively and then summing on both sides, we obtain the normal equations for determination of $a_{12.34}$, $b_{13.24}$, $b_{12.34}$, and $b_{14.23}$ which when substituted to above equation gives the least square regression equation of X_1 on X_2, X_3 and X_4.

The regression equation that uses all the factors which have an influence on the dependent variable can be an extremely useful device for estimating a variable. The chief difficulty, however, with this type of analysis has been the burden of making the calculations. As the number of variables increases, the number of equations to be solved simultaneously and the number of cross products to sum increase to the point that the burden of the arithmetic makes the analysis extremely difficult. The electronic computer is ideally adopted to do this type of analysis and with its help it is possible to use a large number of variables and a large sample and perform all the calculations in a matter of a few seconds.

Relationship between Partial and Multiple Correlation Coefficients

Interesting results connecting the multiple correlation coefficients and the various partial correlation coefficients can be found. For example, we find :

$$1-R^2_{1.23} = (1-r^2_{12})(1-r^2_{13.2})$$

$$1-R^2_{1.234} = (1-r^2_{13})(1-r^2_{13.2})(1-r^2_{14.23})$$

Illustration 5. Find multiple linear regression equation of X_1 on X_2 and X_3 from the data relating to three variables given below :

X_1 :	4	6	7	9	13	15
X_2 :	15	12	8	6	4	8
X_3 :	30	24	20	14	10	4

Solution. The regression equation of X_1 on X_2 and X_3 is

$$X_1 = a_{1.23} + b_{12.3}\,X_2 + b_{13.2}\,X_3$$

The value of the constants $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are obtained by solving the following three normal equations :

$$\Sigma X_1 = Na_{1.23} + b_{12.3}\,\Sigma X_2 + b_{13.2}\,\Sigma X_3$$

$$\Sigma X_1 X_2 = a_{1.23}\,\Sigma X_2 + b_{12.3}\,\Sigma X_2^2 + b_{13.2}\,\Sigma X_2 X_3$$

$$\Sigma X_1 X_3 = a_{1.23}\,\Sigma X_3 + b_{12.3}\,\Sigma X_2 X_3 + b_{13.2}\,\Sigma X_3^2$$

Calculating the required values :

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2	X_1^2
4	15	30	60	120	450	225	900	16
6	12	24	72	144	288	144	576	36
7	8	20	56	140	160	64	400	49
9	6	14	54	126	84	36	196	81
13	4	10	52	130	40	16	100	169
15	3	4	45	60	12	9	16	225
$\Sigma X_1 = 54$	$\Sigma X_2 = 48$	$\Sigma X_3 = 102$	$\Sigma X_1X_2 = 339$	$\Sigma X_1X_3 = 720$	$\Sigma X_2X_3 = 1{,}034$	$\Sigma X_2^2 = 494$	$\Sigma X_3^2 = 2{,}188$	$\Sigma X_1^2 = 576$

Substituting the values in the normal equations :

$$6a_{1.23} + 48b_{12.3} + 102b_{13.2} = 54 \quad \ldots (i)$$

$$48a_{1.23} + 49b_{12.3} + 1034b_{13.2} = 339 \quad \ldots (ii)$$

$$102a_{1.23} + 1034b_{12.3} + 2188b_{13.2} = 720 \quad \ldots (iii)$$

Multiplying Eqn. (*i*) by 8, we get

$$48a_{1.23} + 384b_{12.3} + 816b_{13.2} = 432 \quad \ldots (iv)$$

Subtracting Eqn. (*ii*) from (*iv*), we get

$$110b_{12.3} + 218b_{13.2} = -93 \quad \ldots (v)$$

Multiplying Eqn. (*i*) by 17, we get

$$102a_{1.23} + 816b_{12.3} + 1734b_{13.2} = 918 \quad \ldots (vi)$$

Subtracting Eqn. (*iii*) from Eqn. (*vi*) we get

$$218b_{12.3} + 454b_{13.2} = -198 \quad \ldots (vii)$$

Multiplying Eqn. (*v*) by 109, we obtain

$$11990b_{12.3} + 23762b_{13.2} = -10137 \quad \ldots (viii)$$

Multiplying Eqn, (*vii*) by 55, we get

$$11990b_{12.3} + 24970b_{13.2} = -10890 \quad \ldots (ix)$$

Subtracting Eqn. (*viii*) from Eqn. (*ix*), we get

$$1208b_{13.2} = -753$$

$$b_{13.2} = \frac{-753}{1208} = -0.623.$$

Substituting this value of $b_{13.2}$ in Eqn. (*v*), we get

$$110b_{12.3} + 218(-0.623) = -93$$

$$110b_{12.3} = 135.814 - 93$$

$$b_{12.3} = \frac{42.814}{110} = +0.389.$$

Substituting the value of $b_{12.3}$ and $b_{13.2}$ in Eqn. (*i*), we get

$$6a_{1.23} + 48(0.389) + 102(-0.623) = 54$$

$$6a_{1.23} = 54 + 63.546 - 18.672 = 98.874$$

$$a_{1.23} = 16.479.$$

Thus, the required regression equation is :

$$X_1 = 16.479 + 0.389X_2 - 0.623X_3.$$

Illustration. 6. Given the following, determine the regression equation of :

(*i*) x_1 on x_2 and x_3, and (*ii*) x_2 on x_1 and x_3

$r_{12} = 0.8$ $r_{13} = 0.6$ $r_{23} = 0.5$

$\sigma_1 = 10$ $\sigma_2 = 8$ $\sigma_3 = 5.$

Solution. Regression equation of X_1 on X_2 and X_3 is given by

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3.$$

If the variates X_1, X_2 and X_3 are measured as deviations from their respective means, 'a' will be zero. The values of $b_{12.3}$ and $b_{13.2}$ can be calculated from the data given above but not for 'a'. So, let us assume x_1 and x_2 represent deviations from means. So the regression equation of x_1 on x_2 and x_3 is :

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{13}r_{23}}{1 - r^2_{23}}$$

$$= \frac{10}{8} \times \frac{0.8 - (0.6)(0.5)}{1 - (0.5)^2} = 0.833.$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{13} - r_{13}r_{23}}{1 - r^2_{23}}$$

$$= \frac{10}{5} \times \frac{0.6 - (0.8)(0.5)}{1 - (0.5)^2} = 0.533.$$

$\therefore$ Required regression equation is

$$x_1 = 0.833x_2 + 0.533x_3.$$

(*ii*) Regression equation of x_2 on x_1 and x_3

$$x_2 = b_{21.3}x_1 + b_{23.1}x_3$$

$$b_{21.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{23}r_{13}}{1 - r^2_{13}}$$

$$= \frac{8}{10} \times \frac{(0.8) - (0.5)(0.6)}{1-(0.6)^2}$$

$$= \frac{8}{10} \times \frac{0.8 - 0.3}{1 - 0.36} = \frac{8}{10} \times \frac{05}{064} = 0.625.$$

$$b_{23.1} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{23} - r_{12}\, r_{13}}{1 - r^2_{13}}$$

$$= \frac{8}{5} \times \frac{0.5 - (0.8)(0.6)}{1-(0.6)^2} = \frac{8}{5} \times \frac{0.02}{0.64} = 0.05.$$

Thus $x_2 = 0.625\, x_1 + 0.05\, x_3$ is the required regression equation.

Reliability of Estimates

The problem of determining the accuracy of estimates from the multiple regression is basically the same as for estimates from a simple regression equation. Since the correlation is seldom perfect, estimates made from regression equation will deviate from the correct value or the dependent variable. If an estimate is to be of maximum usefulness, it is necessary to have some indication of its precision. Just as with the simple regression equation, the measure of reliability is an average of the deviation of the actual value of non-dependent variable from the estimates from the regression equation or in other words, the standard error of estimate.

The standard error of estimate of X_1 on X_2 and X_3, is defined as

$$S_{1.23} = \sqrt{\frac{\Sigma (X_1 - X_{1\,est})^2}{N-3}}.$$

$S_{1.23}$ represents standard error of estimate of X_1 on X_2 and X_3, $X_{1\,est}$ indicates the estimated value of X_1 as calculated from the regression equations.

It terms of the correlation coefficients r_{12}, r_{13} and r_{23}, the standard error of estimate can also be computed from the results :

$$S_{1.23} = S_1 \sqrt{\frac{1 - r^2_{12} - r^2_{13} - r^2_{23} + 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}}.$$

MISCELLANEOUS ILLUSTRATIONS

Illustration 7. In a trivariate distribution $\sigma_1 = 3, \sigma_2 = \sigma_3 = 5, r_{12} = 0.6, r_{23} = r_{31} = 0.8$

Find (*i*) $r_{23.1}$, and (*ii*) $R_{1.23}$.

Solution .

(*i*) $$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1 - r^2_{12}}\,\sqrt{1 - r^2_{13}}} = \frac{0.8 - 0.6 \times 0.8}{\sqrt{1-(0.6)^2}\,\sqrt{1-(0.8)^2}} = \frac{0.8 - 0.48}{\sqrt{0.64}\,\sqrt{0.36}} = 0.667.$$

(*ii*) $$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

$$= \sqrt{\frac{(0.6^2 + 0.8^2) - 2(0.6)(0.8)(0.8)}{1-(0.8)^2}} = \sqrt{\frac{(0.36 + 0.64) - 0.768}{0.36}} = 0.803$$

Illustration 8. Calculate (*a*) $R_{1.23}$, (*b*) $R_{3.15}$ and (*c*) $R_{2.13}$ for the following data :

$\overline{X}_1 = 6.8, \overline{X}_2 = 7.0, \overline{X}_3 = 74, S_1 = 1.0, S_2 = 0.8, S_3 = 9, r_{12} = 0.6, r_{13} = 0.7, r_{23} = 65.$

Solution. $$R_{1.23} = \sqrt{\frac{r^2 12 + r^2 13 - 2\, r_{12}\, r_{13}\, r_{23}}{1 - r^2 23}} = \sqrt{\frac{(0.6)^2 + (0.7)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1-(0.65)^2}}$$

$$= \sqrt{\frac{0.36 + 0.49 - 0.546}{0.5775}} = \sqrt{0.527} = 0.726$$

$$R_{3.12} = \sqrt{\frac{r^2_{13} + r^2_{23} - 2 r_{12} r_{13} r_{23}}{1 - r^2_{12}}} = \sqrt{\frac{(0.7)^2 + (0.65)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1 - (0.6)^2}}$$

$$= \sqrt{\frac{0.49 + 0.4225 - 0.546}{1 - 0.36}} = \sqrt{0.573} = 0.757$$

$$R_{2.13} = \sqrt{\frac{r^2_{13} + r^2_{23} - 2 r_{12} r_{13} r_{23}}{1 - r^2_{13}}} = \sqrt{\frac{(0.6)^2 + (0.65)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{0.36 + 0.4225 - 0.546}{0.51}} = \sqrt{0.464} = 0.681.$$

Illustration 9. The following constants are obtained from measurement on length in m.m. (X_1), volume in c.c (X_2) and weight in gm (X_3) of 300 eggs :

$\overline{X}_1 = 55.95$ $S_1 = 2.26$ $r_{12} = 0.578$
$\overline{X}_2 = 51.48$ $S_2 = 4.39$ $r_{13} = 0.581$
$\overline{X}_3 = 56.03$ $S_3 = 4.41$ $r_{23} = 0.974$

Obtain the linear regression equation of egg weight on egg length and egg volume. Hence estimate the weight of an egg whose length is 58 m.m. and volume is 52.5 c.c.

Solution. We have to obtain linear regression equation of egg weight on egg length and egg volume, *i.e.*, X_3 on X_1 and X_2. The regression equation of X_3 on X_1 and X_2 can be written as :

$$X_3 - \overline{X}_3 = \left(\frac{r_{23} - r_{12} r_{13}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_2}\right)(X_2 - \overline{X}_2) + \left(\frac{r_{13} - r_{12} r_{23}}{1 - r^2_{13}}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \overline{X}_1).$$

Substituting the values

$$X_3 - 56.03 = \left(\frac{0.974 - (0.581) \times 0.578}{1 - (0.578)^2}\right)\left(\frac{4.41}{4.39}\right)(X_2 - 51.48) + \left(\frac{0.581 - (0.974 \times 0.578)}{1 - (0.581)^2}\right)\left(\frac{4.41}{2.26}\right)(X_1 - 55.95)$$

$$X_3 - 56.03 = \left(\frac{0.974 - 0.336)}{1 - 0.334}\right)\left(\frac{4.41}{4.39}\right)(X_2 - 51.48) + \left(\frac{0.581 - 0.563}{1 - 0.338}\right)\left(\frac{4.41}{2.26}\right)(X_1 - 55.95)$$

$$X_3 - 56.03 = 0.962\,(X_2 - 51.48) + 0.053\,(X_1 - 55.95)$$

$$X_3 - 56.03 = 0.962X_2 - 49.52 + 0.053X_1 - 2.97$$

$$X_3 = 3.54 + 0.053X_1 + 0.962X_2$$

When length, *i.e.*, X_1 is 58 and volume, *i.e.*, X_2 is 52.5, weight of the egg would be :

$$X_3 = 4.37 + 0.040\,(58) + 0.962\,(52.5) = 4.37 + 2.32 + 50.50 = 57.19 \text{ gm.}$$

Illustration 10. The table shows the corresponding values of three variables X_1, X_2, and X_3. Find the least square regression of X_3 on X_1 and X_2. Estimate X_3 when $X_1 = 10$ and $X_2 = 6$.

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

Solution. The regression equation of X_3 on X_2 and X_1 can be written as follows :

$$X_3 - \overline{X}_3 = \left(\frac{r_{23} - r_{13} r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_2}\right)(X_2 - \overline{X}_2) + \left(\frac{r_{13} - r_{23} r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \overline{X}_1)$$

Calculating $\overline{X}_1, \overline{X}_2, \overline{X}_3, S_1, S_2, S_3, r_{12}, r_{13}, r_{23}$.

X_1	$(X_1-\overline{X}_1)$ x_1	x_1^2	X_2	$(X_2-\overline{X}_2)$ x_2	x_2^2	X_3	$(X_3-\overline{X}_3)$ x_3	x_3^2	x_1x_2	x_1x_3	x_2x_3
3	–5	25	16	+9	81	90	+40	1600	–45	–200	+360
5	–3	9	10	+3	9	72	+22	484	–9	–66	+66
6	–2	4	7	0	0	54	+4	16	0	–8	0
8	0	0	4	–3	9	42	–8	64	0	0	+24
12	+4	16	3	–4	16	30	–20	400	–16	–80	+80
14	+6	36	2	–5	25	12	–38	1444	–30	–228	+190

$\Sigma X_1 = 48$ $\Sigma x_1 = 0$ $\Sigma x_1^2 = 90$ $\Sigma X_2 = 42$ $\Sigma x_2 = 0$ $\Sigma x_2^2 = 140$ $\Sigma X_3 = 300$ $\Sigma x_3 = 0$ $\Sigma x_3^2 = 4008$ $\Sigma x_1x_2 = -100$ $\Sigma x_1x_3 = -582$ $\Sigma x_2x_3 = 720$

$$\overline{X}_1 = \frac{48}{6} = 8, \qquad \overline{X}_2 = \frac{42}{6} = 7, \qquad \overline{X}_3 = \frac{300}{6} = 50$$

$$S_1 = \sqrt{\frac{\Sigma(X_1 - \overline{X}_1)^2}{N}} = \sqrt{\frac{90}{6}} = 3.87$$

$$S_2 = \sqrt{\frac{S(X_2 - \overline{X}_2)^2}{N}} = \sqrt{\frac{140}{6}} = 4.83$$

$$S_3 = \sqrt{\frac{S(X_3 - \overline{X}_3)^2}{N}} = \sqrt{\frac{4008}{6}} = 25.85$$

$$r_{12} = \frac{S\,x_1\,x_2}{\sqrt{S\,x_1^{\,2} \times S\,x_2^{\,2}}} = \frac{-100}{\sqrt{90 \times 140}} = -0.891$$

$$r_{13} = \frac{S\,x_2\,x_3}{\sqrt{S\,x_1^{\,2} \times S\,x_2^{\,3}}} = \frac{-582}{\sqrt{90 \times 4008}} = -0.969$$

$$r_{23} = \frac{S\,x_2\,x_3}{\sqrt{S\,x_2^{\,2} \times S\,x_3^{\,2}}} = \frac{720}{\sqrt{140 \times 4008}} = 0.961$$

$$X_3 - 50 = \left[\frac{0.961 - (-0.969 \times -0.891)}{1 - (-0.891)^2}\right]\left(\frac{25.85}{4.83}\right)(X_2 - 7) + \left(\frac{-0.969 - (0.961 \times -0.891)}{1 - (-0.891)^2}\right)\left(\frac{25.85}{3.87}\right)(X_1 - 8)$$

$$X_3 - 50 = 2.546\,(X_2 - 7) - 3.664\,(X_1 - 8)$$

$$X_3 - 50 = 2.546\,X_2 - 17.822 - 3.664\,X_1 + 29.312$$

$$X_3 = 2.546\,X_2 - 3.664\,X_1 + 61.49.$$

When $X_1 = 10$ and $X_2 = 6$, X_3 will be

$$X_3 = 15.276 - 36.64 + 61.49 = 40.126 \text{ or } 40.$$

Illustration. 11. Suppose a computer has found, for a given set of values of X_1, X_2 and X_3

$r_{12} = 0.96$, $r_{13} = 0.36$ and $r_{23} = 0.78$.

Examine whether these computations may be said to be free from errors.

Solution. For determining whether the given computations are free from errors, or not, we compute the value of $r_{12.3}$. If $r_{12.3}$ comes out to be greater than one, the computations cannot be regarded as free from errors.

$$r_{12.3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{1 - r_{13}^2}\,\sqrt{1 - r_{23}^2}} = \frac{0.96 - 0.36 \times 0.78}{\sqrt{1 - (0.36)^2}\,\sqrt{1 - (0.78)^2}}$$

$$= \frac{0.96 - 0.2808}{\sqrt{0.8704}\,\sqrt{0.3916}} = \frac{0.6792}{0.9329 \times 0.6258} = \frac{0.6792}{0.5838} = 1.163$$

Since $r_{12.3}$ is greater than one, the given computations about r_{12}, r_{13}, etc., do contain some error.

Illustration 12. If $r_{12} = 0.6$, $r_{13} = 0.5$ and $r_{23} = 0.2$, compute the values of $r_{12.3}$ and $R_{1.23}$.

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}$$

$r_{12} = 0.6$, $r_{13} = 5$, $r_{23} = 0.2$. Substituting the values, we get

$$r_{12.3} = \frac{0.6 - 0.5 \times 0.2}{\sqrt{1 - (0.5)^2}\ \sqrt{1 - (0.2)^2}} = \frac{0.6 - 0.1}{\sqrt{0.75 \times 0.96}}\ \frac{0.5}{0.8485} = 0.589$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 r_{12}\, r_{13}\, r_{23}}{! - r_{23}^2}} = \sqrt{\frac{(0.6)^2 + (0.5)^2 - 2(0.6)\,(0.5)\,(0.2)}{1 - (0.2)^2}}$$

$$= \sqrt{\frac{0.36 + 0.25 - 0.12}{1 - 0.04}} = \sqrt{\frac{0.49}{0.96}} = 0.714.$$

Illustration 13. The simple correlation coefficients between variables X_1, X_2 and X_3 are respectively $r_{12} = 0.41$, $r_{13} = 0.71$ and $r_{23} = 0.50$. Calculate the partial correlation coefficients $r_{12.3}$, $r_{23.1}$ and $r_{31.2}$.

Solution. We are given $r_{12} = 0.41$; $r_{13} = 0.71$; $r_{23} = 0.5$. We have to find $r_{12.3}$, $r_{23.1}$ and $r_{31.2}$.

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}} = \frac{0.41 - (0.71 \times 0.5)}{\sqrt{1 - (0.7)^2}\ \sqrt{1 - (0.5)^2}}$$

$$= \frac{0.41 - 0.355}{\sqrt{0.51}\ \sqrt{0.7}} = \frac{0.055}{0.60} = 0.09$$

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{13}^2}} = \frac{0.5 - (0.41 \times 0.71)}{\sqrt{1 - (0.41)^2}\ \sqrt{1 - (0.71)^2}}$$

$$= \frac{0.5 - 0.2911}{\sqrt{0.8319}\ \sqrt{0.4959}} = \frac{0.2089}{0.6423} = 0.325$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{23}^2}} = \frac{0.71 - 0.41 \times 0.5}{\sqrt{1 - (0.41)^2}\ \sqrt{1 - (0.5)^2}}$$

$$= \frac{0.71 - 0.205}{\sqrt{0.8319}\ \sqrt{0.75}} = \frac{0.505}{0.7899} = 0.639$$

Illustration 14. In a trivariate distribution :

$$\bar{X}_1 = 28.20,\ \bar{X}_2 = 4.91,\ \bar{X}_3 = 594,\ S_1 = 4.4,\ S_2 = 1.1,\ S_3 = 80$$
$$r_{12} = 0.80,\ r_{23} = -0.56,\ r_{31} = -0.40.$$

(*a*) Find the correlation coefficient $r_{23.1}$ and $R_{1.23}$.

(*b*) Also estimate the value of X_1, when $X_2 = 6.0$ and $X_3 = 650$.

Solution. (*a*)

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{13}^2}} = \frac{-0.56 - (0.8 \times -0.4)}{\sqrt{1 - (0.8)^2}\ \sqrt{1 - (0.4)^2}}$$

$$= \frac{-0.56 + 0.32}{\sqrt{1 - 0.64}\ \sqrt{1 - 0.16}} = \frac{-0.24}{\sqrt{0.36 \times 0.84}} = -0.436$$

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2 r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2 + (-0.4)^2 - 2(0.8)(-0.4)(-0.56)}{1 - (0.56)^2}}$$

$$= \sqrt{\frac{0.64 + 0.16 - 0.3584}{1 - 0.3136}} = \sqrt{\frac{0.4416}{0.6864}} = +0.802$$

(*b*) The regression equation of X_1 on X_2 and X_3 can be written as follows :

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{S_1}{S_3}\right)(X_3 - \bar{X}_3)$$

$$X_1 - 28.02 = \left(\frac{0.8 - (-0.4) \times (0.56)}{1 - (-0.56)^2}\right)\left(\frac{4.4}{1.1}\right)(X_2 - 4.91) + \left(\frac{(-0.4) - (0.8)(-0.56)}{1 - (-0.56)^2}\right)\left(\frac{4.4}{80}\right)(X_3 - 0.594)$$

$$X_1 - 28.02 = \left(\frac{0.8 - 0.224}{0.6864}\right)\left(\frac{4.4}{1.1}\right)(X_2 - 4.91) + \left(\frac{-0.4 + 0.448}{0.6864}\right)\left(\frac{4.4}{80}\right)(X_3 - 0.594)$$

$$X_1 - 28.02 = 3.356\,(X_2 - 4.91) + 0.070\left(\frac{4.4}{80}\right)(X_3 - 0.594)$$

$$X_1 = 28.02 + 3.356\,X_2 - 16.478 + 0.0039\,(X_3 - 0.594)$$

$$= 11.539 + 3.356\,X_2 + 0.0039X_3$$ is the required regression equation of X_1 on X_2 and X_3.

When $X_2 = 6$ and $X_3 = 650$, estimated value of

$$X_1 = 11.539 + 3.356(6) + 0.0039(650)$$

$$= 11.539 + 20.136 + 2.535 = 34.2107.$$

Illustration 15. The simple correlation coefficients between temperature (X_1), corn yield (X_2) and rainfall (X_3) are : $r_{12} = 0.59$, $r_{13} = 0.46$ and $r_{23} = 0.77$

Calculate partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$.

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{0.59 - (0.46 \times 0.77)}{\sqrt{1 - (0.46)^2}\sqrt{1 - (0.77)^2}}$$

$$= \frac{0.59 - 0.354}{\sqrt{0.7884 \times 0.4071}} = \frac{0.236}{0.5665} = 0.417$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{32}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.59)^2 + (0.46)^2 - 2(0.59)(0.46)(0.77)}{1 - (0.77)^2}} = \sqrt{\frac{0.3481 + 0.2116 - 0.418}{1 - 0.5929}} = \sqrt{\frac{0.1417}{0.4071}} = 0.59.$$

Illustration 16. If $r_{12} = 0.8$, $r_{13} = -0.4$, and $r_{23} = -0.56$, find the value of $r_{12.3}$, $r_{13.2}$, and $r_{23.1}$.

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{0.8 - (-0.4)(-0.56)}{\sqrt{1 - (0.4)^2}\sqrt{1 - (-0.56)^2}}$$

$$= \frac{0.8 - 0.224}{\sqrt{0.84 \times 0.6864}} = \frac{0.576}{0.759} = 0.759$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} = \frac{-0.4 - (0.8)(-0.56)}{\sqrt{1 - (0.8)^2}\sqrt{1 - (-0.56)^2}}$$

$$= \frac{-0.4 + 0.448}{\sqrt{1 - 0.64}\sqrt{1 - 0.3136}} = \frac{0.048}{\sqrt{0.36 \times 0.6864}} = \frac{0.048}{0.4971} = 0.097$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}} = \frac{-0.56 - 0.8 \times -0.4}{\sqrt{1 - (0.8)^2}\sqrt{1 - (0.4)^2}}$$

$$= \frac{-0.56 + 0.32}{\sqrt{1 - 0.64}\sqrt{1 - 0.16}} = \frac{0.24}{0.55} = 0.436$$

Illustration 17. In a trivariate distribution

$r_{12} = 0.863$, $r_{13} = 0.648$ and $r_{23} = 0.709$. Find $r_{12.3}$ and $R_{1.23}$.

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}} = \frac{0.863 - (0.648 \times 0.709)}{\sqrt{1-(0.648)^2}\ \sqrt{1-(0.709)^2}}$$

$$= \frac{0.863 - 0.4594}{\sqrt{0.58 \times 0.50}} = \frac{0.4036}{0.5385} = 0.749$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}} = \sqrt{\frac{(0.863)^2 + (0.648)^2 - 2(0.863)(0.648)(0.709)}{1-(0.709)^2}}$$

$$= \sqrt{\frac{0.7448 + 0.4199 - 0.7930}{1 - 0.5027}} = \sqrt{\frac{0.3717}{0.4973}} = 0.865.$$

Illustration 18. If $r_{12} = 0.80$; $r_{13} = -0.56$ and $r_{23} = -0.40$, then obtain, $r_{12.3}$ and $R_{1.23}$.

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}} = \frac{0.8 - (-0.56)(-0.4)}{\sqrt{1-(-0.56)^2}\ \sqrt{1-(-0.4)^2}}$$

$$= \frac{0.8 - 0.224}{\sqrt{0.6864}\sqrt{0.84}} = \frac{0.576}{0.759} = 0.759$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2 + (-0.56)^2 - 2 \times 0.8 \times (-0.56)(-0.4)}{1-(-0.4)^2}}$$

$$= \sqrt{\frac{0.64 + 0.3136 - 0.3584}{1 - 0.16}} = \sqrt{\frac{0.5952}{0.84}} = 0.842.$$

Illustration 19. Calculate (*a*) $R_{1.23}$, (*b*) $R_{3.12}$ and (*c*) $R_{2.13}$ for the following data :

$r_{12} = 0.6 \qquad r_{13} = 0.7 \quad r_{23} = 0.65$

Solution.

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.7)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1-(0.65)^2}}$$

$$= \sqrt{\frac{0.36 + 0.49 - 0.546}{0.5775}} = \sqrt{0.526} = 0.725$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{(0.7)^2 + (0.65)^2 - 2(0.6 \times 0.7 \times 0.65)}{1-(0.6)^2}}$$

$$= \sqrt{\frac{0.49 + 0.4225 - 0.546}{1 - 0.36}} = \sqrt{0.573} = 0.757$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{13}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.65)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1-(0.7)^2}}$$

$$= \sqrt{\frac{0.36 + 0.4225 - 0.546}{0.51}} = \sqrt{0.464} = 0.681.$$

PROBLEMS

1-A: Answer the following questions, each question carries **one** mark:

(*i*) What is the difference between $r_{12.3}$ and $r_{21.3}$?

(*ii*) Write the formula for partial correlation $r_{23.1}$. *(MBA, Madurai-Kamaraj Univ., 2003)*

(*iii*) Define partial and multiple correlation.

(*iv*) What purpose does partial correlation coefficient serves ?

(*v*) Write the formula for coefficient of multiple correlation $R_{1.23}$.

(*vi*) What is coefficient of determination ?

(*vii*) Write the formula for standard error of estimate $S_{1.23}$.

(*viii*) What is the difference between simple linear and multiple linear regression ?

(*ix*) How multiple correlation differs from partial correlation ?

(*x*) What do you understand by reliability of estimates ?

1-B : Answer the following questions, each question carries **four** marks:

(*i*) In a three variate multiple correlation analysis, the following results were obtained :
$r_{12} = 0.7$, $r_{13} = 0.6$, and $r_{23} = 0.4$
Find the multiple correlation coefficient $R_{1.23}$. *(M. Com., M.K. Univ., 2002)*

(*ii*) Describe the three steps in the process of multiple regression and correlation analysis. *(MBA, Madras Univ., 2003)*

(*iii*) What are zero-order, first-order and second-order coefficients ?

(*iv*) What are the uses and limitations of partial correlation analysis ?

(*v*) What are the advantages and limitations of multiple correlation analysis ?

(*vi*) What are 'normal equations' and how are they used in multiple regression analysis ?

2. Define partial and multiple correlation. With the help of an example distinguish clearly between partial and multiple correlation.
3. What is partial correlation ? Under what circumstances is it to be preferred to the total correlation?
4. (*a*) What is multiple linear regression? Explain clearly the difference between simple linear and multiple linear regression.
   (*b*) With the help of an example illustrate how does multiple linear regression help in the analysis of business problems.
5. Explain the concept of multiple regression and try to find out an example in practical field where multiple regression analysis is likely to be helpful.
6. Distinguish between partial and multiple correlation and point out their usefulness in statistical analysis.
7. Explain the terms : (*i*) Coefficient of determination. (*ii*) Regression coefficient, and (*iii*) Partial and multiple correlation.
8. How do we determine the reliability of estimates obtained from the multiple regression of X_1 on X_2 and X_3 ?
9. (*a*) In the multiple regression equation of X_1 on X_2 and X_3, what are the two regression coefficients and how do you interpret them ?
   (*b*) Explain the concepts of simple, partial and multiple correlation.
   (*c*) When is multiple regression needed? Explain with the help of an example. *(M.Com, DU, 2004)*
10. Within what limits the coefficient of multiple correlation $R_{1.23}$ lies? What inference would you draw if $R_{1.23} = 0$, $R_{1.23} = 1$, $R_{1.23} = 0.92$?
11. How do we distinguish between zero-order, first-order and second-order correlation coefficients ? Illustrate your answer with the help of some examples.
12. What precautions do you think must be observed while making use of partial and multiple correlation techniques ?
13. If $r_{12} = 0.6$, $r_{13} = 0.8$ and $r_{23} = -0.4$, find the values of $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$. Also calculate $R_{1.23}$ and $R_{3.12}$.
14. Calculate $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$ for the following :
    $r_{12} = 0.6$, $r_{13} = 0.7$, $r_{23} = 0.65$
    and comment on these values.
15. In a certain investigation, the following values were obtained :
    $r_{12} = 0.8$, $r_{13} = 0.2$, $r_{23} = -0.5$.
    Do you think that the computations are free from error ?

16. The following information about a trivariate population is given to you :

$$\sigma_1 = 3.2,\ \sigma_2 = 4.5,\ \sigma_3 = 2.8,\ r_{12} = 0.3,\ r_{23} = 6 \text{ and } r_{13} = 0.8.$$

Do you think that the given data are consistent ? If so, calculate $r_{23.1}$ and $r_{1.23}$.

17. Given the following data, find the regression equation of X_1 on X_2 and X_3.

X_1 :	12	22	32	28
X_2 :	6	12	16	22
X_3 :	4	6	12	18

Also predict the value of X_1 when $X_2 = 5$ and $X_3 = 7$.

18. Given the following data :

Performance evaluation	(X_1) :	28	33	21	40	38	46
Aptitude Test Score	(X_2) :	74	87	69	69	81	97
Prior Experience	(X_3) :	5	11	4	9	7	10

(*i*) Develop the estimating equation, best describing these data.

(*ii*) If an employee scored 83 on the aptitude test and had a prior experience of 7 years, what performance evaluation would be expected ?

(M.Com., DU, 2007)

*****

Useful Books for
M.B.A., P.G. Diplomas in Management and M.Com. of various Universities and Management Institutes

Principles and Practice of Management	Dr. L.M. Prasad
Management Theory and Practice	Dr. C.B. Gupta
Management Process and Organisational Behaviour	Dr. L.M. Prasad
Organisational Behaviour	Dr. L.M. Prasad
Organisation Development & Human Resource Development	Dr. P.C. Tripathi
Human Resource Development	Dr. P.C. Tripathi
Personnel Management and Industrial Relations	Dr. P.C. Tripathi
Human Resource Management	Dr. L.M. Prasad
Human Resource Management	Dr. C.B. Gupta
Strategic Planning & Management	Dr. P.K. Ghosh
Business Policy & Strategic Management	Dr. L.M. Prasad
Corporate Planning & Policy	Dr. C.B. Gupta
Financial & Management Accounting	Dr. S.N. Maheshwari
Financial & Cost Accounting	Dr. S.N. Maheshwari
Management Accounting—Text	T.S. Grewal, Dr. Hingorani & A.R. Ramanathan
Management Accounting & Financial Control	Dr. S.N. Maheshwari
Principles & Practice of Financial Management	Dr. S.N. Maheshwari
Investment Management Theory & Practice	R.P. Rustagi
Financial Analysis and Financial Management	Dr. R.P. Rustagi
Indian Financial System	Dr. P.N. Varshney & Dr. D.K. Mittal
International Financial Management	Dr. R.L. Varshney & S. Bhashyam
Foreign Exchange & Risk Management	C. Jeevanandam
Managerial Economics	Dr. R.L. Varshney & Dr. K.L. Maheshwari
Managerial Economics	Dr. P.L. Mehta
Economic Environment of Business	Dr. M. Adhikary
Indian Economy—Environment & Policy	I.C. Dhingra
Business Environment	C.B. Gupta
Business Mathematics	Dr. D.C. Sancheti & V.K. Kapoor
Business Statistics	Dr. S.P. Gupta & Dr. M.P. Gupta
Operations Research—Techniques for Management	V.K. Kapoor
Operations Research	Prof. Kanti Swaroop, Dr. P.K. Gupta & Dr. Man Mohan
Operations Research—Problems & Solutions	V.K. Kapoor
Problems in Operations Research	Dr. P.K. Gupta & Dr. Man Mohan
Elements of Mercantile Law	N.D. Kapoor
Legal and Regulatory Framework of Business	N.D. Kapoor
Essentials of Business Communication	Korlahalli & Rajendra Pal
Marketing Management	Dr. R.L. Varshney & Dr. S.L. Gupta
Marketing Management	C.B. Gupta & Rajan Nair
International Marketing Management	Dr. R.L. Varshney & Prof. Bhattacharya
Marketing Research	Dr. D.D. Sharma
Consumer Behaviour	Dr. S.L. Gupta & Sumitra Pal
Advertising & Sales Promotion	Dr. S.L. Gupta and Dr. V.V. Ratna
Textbook of Research Methodology in Social Sciences	Dr. P.C. Tripathi
Materials Management	Dr. M.M. Varma
Project Management & Control	Dr. P.C.K. Rao
Management Information Systems	Dr. L.M. Prasad & Usha Prasad
Computers and Information Technology	V.K. Kapoor
The Pinnacle of Success	Dr. D.D. Sharma
Entrepreneurship Development in India	Dr. C.B. Gupta & Dr. N.P. Srinivasan
Entrepreneurship & Small Business Management	Dr. C.B. Gupta & Dr. S.J. Khanka
Steps to Success	Mahatma Devesh Bhikshu
Total Quality Management	Dr. D.D. Sharma

Sultan Chand & Sons

23, Daryaganj, New Delhi-110 002
Phone : 23243183, 23247051, 23266105, 23277843, 23281876, Fax: 011-23266357

ISBN 81-8054-641-9

TC 518

21

Statistical Decision Theory

INTRODUCTION

Each one of us is concerned with making decisions of one kind or the other. Some of these are really difficult to make because of the complexity of a decision situation. It is precisely the complexity that has led many decision-makers to analyse the process of decision-making. It is in this context, that we need a decision theory which may be defined as a body of methods helpful to a decision-maker to select wisely, the best course of action from amongst several alternatives. The problem of statistical decision theory is that, given a situation where there are several available alternative courses of action, each of which may lead to a set of mutually exclusive outcomes associated with certain probabilities, which course of action should a decision-maker take? Any business unit may be faced with various types of problem situations such as: how much product to be produced; how much promotional effort to be expanded; how much quantity to be stocked, etc.

There are some elements which are common to all kinds of decision problems. These are :

1. The Decision-maker. The decision maker refers to an individual or a group of individuals responsible for making a choice of an appropriate course of action from the available courses of action.

2. Courses of Action. Courses of action are also called actions, alternatives, acts or strategies. For a problem situation, all possible courses of action should be included.

3. States of Nature. States of nature are sometimes called outcomes or events. The decision-maker must develop an exhaustive list of possible *future* of events. However, decision-maker has no control over the occurrence of specific event.

4. Payoff. It is the effectiveness associated with a particular combination of a course of action and state of nature.

5. Payoff Table. For a given problem, a payoff table lists the states of nature and a set of given courses of action. For each combination of states of nature and course of action, the payoff is calculated. Let different states of nature or outcomes be represented by $O_1, O_2, ..O_m$ and different courses of actions or strategies by $S_1, S_2, ... S_n$ For a given combination of O_i and S_j, the corresponding payoff is denoted by a_{ij}. Such a payoff table is shown below :

	Courses of actions			
States of nature	S_1	S_2	S_j	S_n
O_1	a_{11}	a_{12}	a_{1j}	a_{1n}
O_2	a_{21}	a_{22}	a_{2j}	a_{2n}
O_i	a_{i1}	a_{i2}	a_{ij}	a_{in}
O_m	a_{m1}	a_{m2}	a_{mj}	a_{mn}

Payoffs may be evaluated in terms of *Profit, Cost,* or *Opportunity Loss.* When such payoffs are evaluated in terms of profit, they are called *Payoffs.* A payoff table shows the relation between all possible outcomes (states of nature), all possible courses of action and the value associated with these.

In the payoff table, the column headings designate the various courses of action out of which the decision-maker may choose while the row headings show the admissible states of nature under which the decision-maker has to take decision. The cell value a_{ij} shows payoff resulting by taking course of action S_j when the state of nature is O_i for all $j = 1, 2, 3, ..., n$ and $i = 1, 2, 3, ..., m$. A payoff table represents the economics of a problem —a problem of revenue and costs. A payoff may be thought of as a conditional value or conditional profits (losses). It is conditional value in the sense that associated with each course of action there is a certain profit (or loss), given that a specific state of nature has occurred. A payoff table thus contains all conditional values of all possible combinations of courses of action and states of nature.

Decision theory models can be classified on the basis of the degree of certainty. We can visualize decision problems with complete certainty on the one hand and complete uncertainty on the other. Therefore, most of the decision situations fall between these two extremes. We can classify the following types of decision-making and each will be discussed in detail :

(*a*) Decision-making under certainty.

(*b*) Decision-making under risk.

(*c*) Decision-making under uncertainty,

(*d*) Decision-making under conflict or theory of games.

(a) Decision-making under Certainty

In this type of deterministic situation, the outcome of a specified decision can be predetermined with certainty.

(b) Decision-making under Risk

There are certain types of decision problems where there can be more than one outcome and it is possible to assign a probability value to each outcome. In other words, the decision-maker knows the probability of occurrence of each state of nature.

(*i*) *Expected monetary value.* The best strategy is selected on the basis of the highest expected monetary value (EMV). The EMV for a course of action is the sum of the products obtained by multiplying the payoff for a given outcome by its probability value.

If $m_1, m_2, ...m_n$ are the payoffs corresponding to the states of nature $S_1, S_2, ...S_n$ respectively, and the corresponding probabilities of $S_1, S_2, ...S_n$ are $P_1, P_2, ... P_n$, then EMV is defined as :

$$\text{EMV} = m_1p_1 + m_2p_2 + ...m_np_n = \Sigma mp.$$

The following points in relation to decision-making under risk may be found relevant :

(*a*) We must be able to construct conditional payoff table in order to compare the EMVs of different courses of action.

(*b*) The choice of the strategy is purely on the basis of EMV. Sometimes, utility may be more appropriate than EMV.

(*c*) The EMV criterion does not take into account the quality of risk. Generally, the decision-maker goes by the trade-off between risk and return.

(*ii*) *Expected opportunity loss.* Another decision criterion of decision theory is called expected opportunity loss (EOL). Opportunity loss represents the amount of profit that is lost because the most profitable course of action is not taken. To calculate EOL, we must find the conditional opportunity loss

(COL). The COL of the optimal act being zero, the COL of any other act is the difference between the payoff of the optimal act and the action taken, and obviously will always be positive. When payoffs are replaced by their corresponding opportunity losses, we get what is known as the Loss Table. If l_{ij} (the element of a loss table) is the opportunity loss resulted by taking an action Sj, when the state of nature O_i then l_{ij} satisfies the following relation :

$$l_{ij} = \max_j p_{ij}\, l - p_{ij}$$

for all $i = 1, 2, \ldots\ldots, n$

$j = 1, 2, \ldots\ldots, m$

Illustration 1. A baker produces a certain type of special pastry at a total average cost of Rs. 3 and sell it at a price of Rs. 5. This pastry is produced over the weekend and is sold during the following week; such pastry being produced but not sold during a week's time are totally spoiled and have to be thrown. According to past experience, the weekly demand for these pastries is never less than 78 or greater than 80. You are required to formulate action space, state space, payoff table and loss table.

Solution. It is clear from the problem given that the manufacturer will not produce less than 78 or more than 80 pastries. Thus, there are three courses of action open to him :

S_1 = produce 78 pastries

S_2 = „ 79 „

S_3 = „ 80 „

The state of nature is the weekly demand for pastries. There are three possible states of nature, *i.e.*,

O_1 = demand is 78 pastries

O_2 = „ 79 „

O_3 = „ 80 „

The uncertainty element in the problem is the weekly demand. The bakery profits are conditioned by the weekly demand. Cell values of payoff table are computed as follows :

a_{11} = payoff when action S_1 is taken but the state of nature is O_1
= Rs. $[5 \times 78 - 3 \times 78]$ = Rs. 156.

a_{12} = payoff when action S_2 is taken but the state of nature is O_1
= Rs. $[5 \times 78 - 3 \times 79]$ = Rs. 153.

a_{13} = payoff when action S_3 is taken but the state of nature is O_1
= Rs. $[5 \times 78 - 3 \times 80]$ = Rs. 150.

Similarly, a_{21} = payoff when action S_1 is taken but the state of nature is O_2
= Rs. $[5 \times 78 - 3 \times 78]$ = Rs. 156.

a_{22} = Rs. $[5 \times 79 - 3 \times 79]$ = Rs. 158.

a_{23} = Rs. $[5 \times 79 - 3 \times 80]$ = Rs. 155.

Similarly, a_{31} = payoff when action S_1 is taken but the state of nature is O_3
= Rs. $[5 \times 78 - 3 \times 78]$ = Rs. 156.

a_{32} = Rs. $[5 \times 79 - 3 \times 79]$ = Rs. 158.

a_{33} = Rs. $[5 \times 80 - 3 \times 80]$ = Rs. 160.

These values are tabulated below :

PAYOFF TABLE

State of nature \ *Courses of Action*	S_1	S_2	S_3
O_1	156	156	150
O_2	158	156	155
O_3	156	158	160

To calculate opportunity losses, we first calculate max a_{1k}, max a_{2k}, and max a_{3k}.

$$\max a_{1k} = 156, \max a_{2k} = 158, \max a_{3k} = 160$$

$$l_{11} = 156 - 156 = 0, l_{12} = 156 - 153 = 3, l_{13} = 156 - 150 = 6$$

$$l_{21} = 158 - 156 = 2, l_{22} = 158 - 158 = 0, l_{23} = 158 - 153 = 5$$

$$l_{31} = 160 - 156 = 4, l_{32} = 160 - 158 = 2, l_{33} = 130 - 130 = 0$$

The loss table corresponding to payoff table is given below :

LOSS TABLE

State of nature / Courses of Action	S_1	S_2	S_3
O_1	0	3	6
O_2	2	0	5
O_3	4	2	0

(iii) Expected value of perfect information. The expected value of perfect information (EVPI) is the difference between expected profit (EP) of the optimal decision without perfect information and that with the perfect information. This expected profit with perfect information (EPPI) is called the expected value of payoff under certainty. The perfect prediction reduces the opportunity losses due to uncertainty to zero. The highest payoff in the absence of perfect predictor is *EP* of the optimal action. The difference between *EPPI* and *EP* is called the *expected value of perfect information* (abbreviated *EVPI*). *EVPI* represents the maximum amount of money which a decision-maker could spend to obtain additional information regarding the states of nature.

It may be noted that *EVPI* is always equal to the *EOL* of selecting the optimum action under uncertainty. The identity $EP + EOL = EPPI$ follows from the result $EVPI = EOL$ and $EVPI = EPPI - EP$.

The main objective of preposterior analysis is to determine whether or not it is profitable to gather additional information regarding the states of nature before taking the final action. Additional information may be gathered by conducting survey, by carrying out an experiment or by some other means. The objective of preposterior analysis is fulfilled by computing *EVPI*. If *EVPI* is relatively larger than the cost involved in gathering additional information, it is advisable to gather the additional information regarding the states of nature.

Illustration 2. A businessman wants to construct a hotel. He usually builds 25,50 or 100 bed hotel, depending on whether anticipated demand is low, medium or high. The businessman has been able to find out net profits which are expressed in the table below and the corresponding probabilities are also given below.

States of nature / Courses of Action	S_1 Build 25-bed-hotel	S_2 Build 50-bed-hotel	S_3 Build 100-bed-hotel
O_1 = Low demand	20,000	–10,000	–30,000
O_2 = Medium demand	25,000	25,000	–5,000
O_3 = High demand	30,000	50,000	60,000

States of nature = Demand	O_1	O_2	O_3	Total
Probability	0.2	0.3	0.5	1.00

(*a*) Compute *EP, EPPI* and *EVPI.*

(*b*) A research firm agrees to conduct a survey for Rs. 8000, and provide him with information regarding the states of nature. Should the survey be conducted ?

Solution. *To compute EP,* we have to compute expected payoff each action under uncertainty as follows :

$$E(S_1) = \text{Rs. } [20{,}000 \times 0.2 + 25{,}000 \times 0.3 + 30{,}000 \times 0.5] = \text{Rs. } 26{,}500$$

$$E(S_2) = \text{Rs. } [-10{,}000 \times 0.2 + 30{,}000 \times 0.3 + 50{,}000 \times 0.5] = \text{Rs. } 32{,}000$$

$$E(S_3) = \text{Rs. } [-30{,}000 \times 0.3 + (-5{,}000) \times 0.3 + 60{,}000 \times 0.5] = \text{Rs. } 19{,}500$$

From the above computation it is clear that the highest expected payoff or profit is associated with action S_2. Hence, the highest expected payoff under certainty = EP = Rs. 32,000.

Next, to compute *EPPI* we have to find out the highest payoff for each action under certainty, *i.e.*, under the assumption that the perfect predictor is available. Clearly from the payoff table, when the state of nature is known to the businessman would take action S_1 as a result of nature are known O_1 to be O_2 and O_3 he correspondingly takes action S_2 and S_3 by which he makes his net profit 30,000 and 60,000 respectively. The highest expected pay off under certainty is computed as $EPPI$ = Rs. [20,000 × 0.2 + 30,000 × 0.3 + 60,000 × 0.5] = Rs. 43,000. The expected value of perfect information = $EVPI = EPPI - EP$ = Rs. [43,000 – 32,000] = Rs. 11,000. The *EVPI* is relatively larger than the expenditure incurred in conducting a survey in order to collect further information regarding the states of nature. Hence, it is advisable to conduct the survey.

(c) Decision-making under Uncertainty

Competitive decision model is one related to the situation of uncertainty, the probabilities of occurrence of the different events (or the states of nature) are not known and the decision-maker has no way of calculating the expected payoffs for his strategies. This, in other words, means that the decision-maker has to act with imperfect information in such a situation. Consequently, there is no single best criterion for selecting a strategy to deal with such a situation but there are different criteria available for selecting a strategy. The following criteria are important in this context :

(*i*) *Maximin decision rule.* Under this rule, the decision-maker is completely pessimistic. He assumes that the situation will always be disadvantageous. As such he selects that strategy which give largest minimum payoffs, *i.e.*, maximum of the minimums.

(*ii*) *Maximax decision rule.* Under this rule, the decision-maker is quite optimistic. He assumes that the situation will always be to his advantage. He, therefore, selects the strategy which yields him the highest possible pay-off, *i.e.*, maximum of the maximums.

(*iii*) *Minimax decision rule.* This rule is based on general insurance against risk. It insures against the maximum possible risk. Under it, one adopts the strategy which causes minimum of the maximum losses. Because of such an attitude, this rule is sometimes also known as "regret rule", for one looks at loss opportunities (losses) as regrets. The minimax rule items from the work of John Von Neumann and Oskar Morgenstern. This rule states that the decision-maker should minimize maximum harm.

(*iv*) *Hurwicz decision rule.* Hurwicz has developed a decision rule basing it on the maximin and maximax rules with an index of optimism (x) and an index of pessimism ($1 - x$). The value of x always lies between zero and one. The decision-maker should assign a value to x somewhere between 0 and 1. The value of x nearer to 1 means the decision-maker is optimistic, and near to 0 reflects a pessimistic

decision-maker and $x = ½$ reflects a neutralist. Largest and smallest values, say V, U respectively be determined for each and every strategy by applying maximin and maximax rules and then the expected value be determined as under :

Expected value $= x.V + (1 - x)\ U$

Then the strategy having the highest expected value as per the above formula given by Hurwicz is selected.

(*v*) *Laplace decision rule.* This rule is based on the assumption (in case of the probabilities are not known) that the probabilities of different states of nature for a given strategy are all equal. Considering these equal probabilities, the expected payoffs will be calculated as per the method above stated and then the strategy with the largest expected payoff is selected.

From the above description, it should become clear that there is no single best rule for decision-making under the situation of uncertainty. There are several models for the purpose. The choice for the selection of a model should be left to the decision-maker who should ultimately decide as per his own skill and experience considering the environment, firm's policy and other relevant factors.

The following example can be illustrated to exhibit how to make decision under uncertainty :

Illustration 3. A Toy Company is bringing out a new type of toy. The company is attempting to decide whether to bring out a full, partial, or minimal product line. The company has three levels of product acceptance. Management will make its decision on the basis of expected profit from the first year of production. The relevant data are shown in the following table :

	Anticipated 1st year Profit (Rs. 000's)		
Product	*Product line*		
Acceptance	*Full*	*Partial*	*Minimal*
Good	80	70	50
Fair	50	45	40
Poor	–25	–10	0

Take optimal decision under each of the following decision criteria :

(*i*) Maximax, (*ii*) Maximin,

(*iii*) Laplace criteria, (*iv*) Minimax regret.

Solution. (*i*) The maximum profit (in Rs. 000's) for each product line is

Full — 80
partial — 70
Minimal — 50

∴ The maximax of maximum payoffs is for full product line Rs. 80,000.

∴ The company should go for full product line under optimistic or maximax criteria.

(*ii*) The minimum profit (in Rs. 000's) for each product line is

Full — –25
partial — –10
Minimal — – 0

∴ The maximum of minimum payoffs is for minimal product line Rs. 0.

∴ The Company should go for minimal product line under pessimistic or maximin criteria.

(*iii*) If chances of Good, Fair or Poor product acceptance are equal *i.e.*, $\frac{1}{3}$, then the expected profit for each product line is

$$\text{Full} = \frac{1}{3}(80{,}000) + \frac{1}{3}(50{,}000) + \frac{1}{3}(-25{,}000)$$

$$= \frac{1}{3}(1{,}05{,}000) = \text{Rs. } 35{,}000$$

$$\text{Partial} = \frac{1}{3}(70.000) + \frac{1}{3}(45{,}000) + \frac{1}{3}(-10{,}000)$$

$$= \frac{1}{3}(1{,}05{,}000) = \text{Rs. } 35.000$$

$$\text{Minimal} = \frac{1}{3}(50{,}000) + \frac{1}{3}(40{,}000) + \frac{1}{3}(0)$$
$$= \text{Rs. } 30{,}000$$

∴ The maximum profit is for Full and Partial product line.

∴ The Company should go for full or partial product line under Laplace criteria.

(*iv*) The conditional opportunity loss table is

Product Acceptance	*Anticipated 1st year Profit (Rs. 000's)*		
	Product line		
	Full	*Partial*	*Minimal*
Good	0	10	30
Fair	0	5	10
Poor	25	10	0
Maximum loss	25	10	30

∴ The minimum of maximum loss is for partial product line.

∴ The company should go for partial product line under minimax regret criteria.

(*d*) Decision-making under Conflict (Theory of Games)

The theory of games which is also called decision-making under conflict, dates back to 1944 with the classic work of J. von Neumann and O. Morgenstern entitled *Theory of Games and Economic Behaviour.* Game theory provides a framework for analysing competitive situations in which the competitors (or players) make use of logical processes and techniques in order to determine an optimal strategy for "winning". Since many situations in business involve competition, game theory is of considerable theoretical interest.

A game can be played between two or more individuals or groups of individuals. Business environment being always competitive, the number of problems which lend themselves to this theory are abundant. For example, two firms may be trying to determine how much of advertising to do, where the options in terms of amounts of advertising, and the payoffs (may be in terms of increase in sales or market share) are known.

A *Game* is described by its set of rules. These rules specify clearly what each person, called a player, is allowed or required to do under all possible sets of circumstances. The rules also define the amount of information, if any, each person receives. A game is finite when each player has a finite number of moves and finite number of choices at each move.

Kinds of Games. It is convenient to classify games according to the number of players, *i.e.*, as two persons, three persons, etc. It is also convenient to distinguish between games whose payoffs are zero sum and those whose are not. If the players make payments only to each other, *i.e.*, the loss of one is the gain of the other, the game is said to be *zero sum.* Thus, solitaire is a one-person game and chess is a two-person game. Mathematically speaking, a zero sum game can be represented as:

In an n person game with players $PL_1, PL_2 \ldots, PL_n$ and payoffs PF_i $(i = 1, 2, \ldots n)$ be made to PL_i at the end of the game (PF_i will be negative if PL_i has to pay rather than receive).

Then, if $\Sigma PEi = 0$, the game is *zero-sum.*

$\neq 0$, the game is non-*zero-sum.*

The usual distinction in game theory is between two-person games and games involving three or more persons. The theory of games of three or more persons (n-person games) is largely undeveloped, and it is precisely this limitation that has restricted the application of game theory from many real life applications. Therefore, the discussion of this chapter will be limited to the presentation and analysis of two-person zero-sum games. The underlying assumptions, the rules of the game are given as follows :

1. The players act rationally and intelligently.
2. Each player has available to him a finite set of possible courses of action.
3. The players attempt to maximize gains and minimize losses.
4. All relevant information is known to each player.
5. The players make individual decisions without direct communication.
6. The players simultaneously select their respective courses of action.

Two-Person Zero-Sum Game

Two-person zero-sum games are the games played by two persons, parties, or groups, with directly opposite interest. One persons's gain in the game is exactly equal to another person's loss, and therefore, the sum total of the gains and lossses equals zero. Each person has alternative choices of *strategies* (moves) available to him, and the rules governing choices are known in advance to the players. The outcome of a set of possible choices of strategies is also known to the two players in advance and is expressed in terms of numerical values.

A two-person zero-sum game is conveniently represented by a game matrix as shown below. A game matrix is often referred to as a *payoff matrix,* because the outcome of the alternative choice of strategies are expressed in terms of payoff units.

		Player B's strategies	
		B_1	B_2.....B_n
Player A's Strategies	A_1	a_{11}	a_{12}.....a_{1n}
	A_2	a_{21}	a_{22}.....a_{2n}
	⋮	⋮	⋮
	A_m	a_{m1}	a_{m2}.....a_{mn}

For example, payoff a_{12} refers to the strategy A_1 adopted by player A and strategy B_1 by player B.

The above payoff matrix is in relation to player A. It is important to note that only player A's gains are included in the payoff matrix. However, if the payoff matrix represents gains by player A, then player B in turn loses the same amount gained by player A and the sum of the reward is zero (zero-sum).

A game with a Pure Strategy

Let us consider a game with a payoff matrix presented in the following table. This is a two person zero-sum game with two alternative choices of strategies available to player A and three alternative choices of strategies available to player B. If the payoffs matrix represents the per cent market share obtained by player A, then player B loses the market share that A gains.

		Player B		
		B_1	B_2	B_3
Player A	A_1	80	40	75
	A_2	70	35	30

For example, if player A selects strategy A_1 and player B selects strategy B_1, then player A *wins* 80 per cent of the market share while player B loses 80 per cent of the market share.

In this game, player A's objective is to select a strategy which enables him to gain as much as possible. In contrast, player B's objective is to select a strategy which enables him to lose as little as possible. If a single strategy is chosen by players A and B, then it is referred to as a *pure strategy*. Therefore, a pure strategy is a single strategy in a stable solution. If the solution is not stable, then, we cannot have a pure strategy without further analysis. The first step in the pure strategies is to determine the minimax and maximin value. If the two values coincide (equal), we get the pure strategies for both the players to adopt. This equal value will be termed as *saddle point* and this value will be the value of the game.

Illustration 4. Consider the two-person zero-sum game with the following payoff matrix :

		Player B		
		B_1	B_2	B_3
Player A	A_1	5	4	6
	A_2	2	3	7
	A_3	4	3	0

Determine the optimal pure strategies for both the players and find the value of the game.

Solution. Using maximin principle, player A selects that strategy which is maximum of the minimum gains (payoffs), *i.e.*, best of the worst guaranteed gains. Similarly, player B selects that strategy which is minimum of the maximum losses (payoffs), *i.e.*, the best of the worst losses. In fact, if the payoffs matrix contains both gains and losses for each player, either criterion will yield the same result. Minimax and maximin, both select the best of the worst outcomes.

Using the maximum principle, the strategy to be chosen will be determined based on the values of row minima. Similarly, for minimax principle (for opponent will be determined based on the values of the column maxima. This is shown below:

		Player B			
		B_1	B_2	B_3	Row minimum
Player A	A_1	5	4	6	4*
	A_2	2	3	7	2
	A_3	4	3	0	0
Column maximum		5	4*	7	

In this game, player A *will choose strategy* A_1, which yields the minimum pay off 4. Similarly, the best strategy for player B is a strategy which lead to a minimum column maxima. In this case, player B will choose strategy B_2 which has a maximum loss of 4. Both the maximum value of row minima and the minimum value of column maxima are denoted by asterisks in the game matrix. Since the value of the maximin coincides with the value of the minimax, *an equilibrium* or *saddle point* is determined in this game. It is apparent that a saddle point is that point which is minimum in the row and maximum in the column. The amount of payoff at an equilibrium point is also known as the *value of the game*. Hence, the optimal pure strategies for both the players are : Player A must select strategy A_1 and player B must select strategy B_2. The value of the game is 4 which indicates that player A will gain 4 units and player B will lose 4 units.

A Game with a Mixed Strategy

With no pure strategy solution, both players will prefer to alter the strategy selection or play *a mixed strategy*. Consider the payoff matrix as below :

		Player B	
		B_1	B_2
Player A	A_1	65	45
	A_2	50	55

Since this game has no saddle point, therefore, we cannot have pure strategies. In such cases, each player would like to mix up his strategies in a random selection. The random selection plan involves

selecting each strategy a certain per cent of the time, such that the player's *expected* gains (or losses) are equal, regardless of the opponent's selection of strategies. Selection of a strategy, a given per cent of the time is analogous to the selection of a strategy with a given probability. There are various methods to solve such type of games, but only three methods will be discussed here :

Method 1 (Algebraic)

The method for determining the per cent (or probability) to be associated with a given strategy wiil be illustrated for the previous example. Let us begin with player A, he wishes to select strategy A_1 or A_2 according to probabilities such that his expected gains are the same, regardless of player B's selection of strategies B_1 or B_2. If player B selects strategy B_1, the possible payoffs to player A are 65 and 50. If player A selects strategy A_1 with a probability of p and, therefore, selects strategy A_2 with a probability of $(1-p)$, then his expected gains for this game are given by :

$$65p + 50\,(1-p)$$

On the other hand, if player B selects strategy B_2, then player A's expected gains are :

$$45p + 55\,(1-p)$$

Now, in order for player A to be indifferent to which strategy player B selects, he wishes his expected gains to be equal for each of player B's possible moves. Thus, the two equations of expected gains are set equal and solved for p as given below :

$$65p + 50\,(1-p) = 45p + 55\,(1-p)$$

or $$25p = 5$$

Therefore, $$p = 1/5 = 0.2\ ;\ 1-p = 1-0.2 = 0.8.$$

Hence, player A would select strategy A_1 with a probability of 0.2 and strategy A_2 with a probability of 0.8.

Similarly, player B would determine his probabilities q and $(1-q)$ for selecting strategies B_1 and B_2 respectively, by equating his expected losses if player A chooses strategy A_1 to the expected losses if player A chooses strategy A_2 as follows :

$$65q + 45\,(1-q) = 50q + 55\,(1-q)$$

or $$25q = 10$$

Therefore, $$q = 2/5 = 0.4\ ;\ 1-q = 1-0.4 = 0.6.$$

Hence, player B would select strategy B_1 with a probability of 0.4 and strategy B_2 with a probability of 0.6.

The value of the game is determined by substituting the value of p or q in any of the expected value and is calculated as 53.

Method 2 (Calculus Method)

This method is almost similar to the previous method except that instead of equating the two expected values, the expected value for a given player is maximised. To illustrate this method, let us take the same example discussed in the previous method.

Suppose player A selects strategy A_1 with a probability p and obviously selects A_2 with a probability $(1-p)$ and player B selects strategy B_2 with a probability q and obviously selects strategy B_2 with a probability $(1-q)$. Then the expectation is given as below :

$$E\,(p, q) = 65pq + 45p\,(1-q) + 50\,(1-p)\,q + 55\,(1-p)\,(1-q)$$

If expectation is to be maximized, then

$$\frac{\partial E}{\partial p} = \frac{\partial E}{\partial q} = 0$$

$$\frac{\partial E}{\partial p} = 65q + 45(1-q) - 50q - 55(1-q) = 0$$

or $25q = 10$

Therefore, $q = \frac{10}{25} = 0.4,\ 1 - q = 1 - 0.4 = 0.6$

and $$\frac{\partial E}{\partial q} = 65p + 45p + 50(1-p) - 55(1-q) = 0$$

or $25p = 5$

Therefore, $p = \frac{5}{25} = 0.2\ ;\ 1 - p = 1 - 0.2 = 0.8$

To determine the value of the game, substitute the values of p, $1-p$, q and $1-q$ in the expression of expected value. The value of game is found to be 53 as before.

Illustration 5. Consider a rectangular game whose matrix is :

		Player B	
		1	2
Player A	1	1	3
	2	4	2

Find the best strategies and the value of the game.

Solution. Since the matrix has no saddle point, it is desirable for A and B to play with certain frequencies. suppose A plays 1 with frequency $x(0 \le x \le 1)$ and plays 2 with frequency $(1-x)$; and suppose B plays 1 with frequency $y\ (0 \le y \le 1)$ and plays 2 with frequency $(1-y)$. Then, A's mathematical expectation is given by

$$E(x, y) = 1xy + 3x(1-y) + 4y(1-x) + 2(1-x)(1-y)$$

If expectation is to be maximised :

$$\frac{\partial E}{\partial x} = y + 3(1-y) - 4y - 2(1-y) = 0 \qquad ...(1)$$

$$\frac{\partial E}{\partial y} = x + 3x + 4(1-x) - (1-x) = 0 \qquad ...(2)$$

From (1), $y + 3 - 3y - 4y - 2 + 2y = 0$

$4y = 1$ or $y = 1/4$

From (2), $x + 3x + 2 - 4x - 2 + 2x = 0$

$4x = 2$ or $x = 1/2$

Thus, here A should choose strategies 1 and 2 with equal probability and B should choose 1 with a probability 1/4 and choose 2 with a probability 3/4. The value of the game is found to be 5/2 by substituting these values of x and y in $E(x, y)$.

Illustration 6. Consider a game whose matrix is given as :

		Player B	
		1	2
Player A	1	a	b
	2	c	d

Find the optimal strategies for each player and the value of the game.

Solution. Let x and y be the probabilities of adopting strategies 1 and 2 by player A and B, respectively.

Then $$E(x,y) = axy + bx(1-y) + cy(1-x) + d(1-x)(1-y) \qquad ...(i)$$

$$\frac{\partial E}{\partial x} = ay + (b - by) - cy - d(1-y) = 0$$

$$y(a - b - c + d) = d - b$$

$$y = \frac{(d-b)}{(a-b-c+d)}$$

$$\frac{\partial E}{\partial y} = ax - bx + c - cx - d(1-x) = 0$$

$$x(a-b-c+d) = d-c$$

$$x = \frac{(d-c)}{(a-b-c+d)}$$

Substituting in (*i*), we get

$$V = \frac{(ab-bc)}{(a+d)-(b+c)}$$

However, this method of solution can only be used for 2 × 2 matrices ; for larger matrices we need to develop some other solution procedure.

Method 3 (Graphical Method)

The graphical method can be used to solve games where one of the players has only two alternatives and the other has two or more alternatives. Consider the following game :

		Player B	
		B_1	B_2
Player A	A_1	1	7
	A_2	6	2

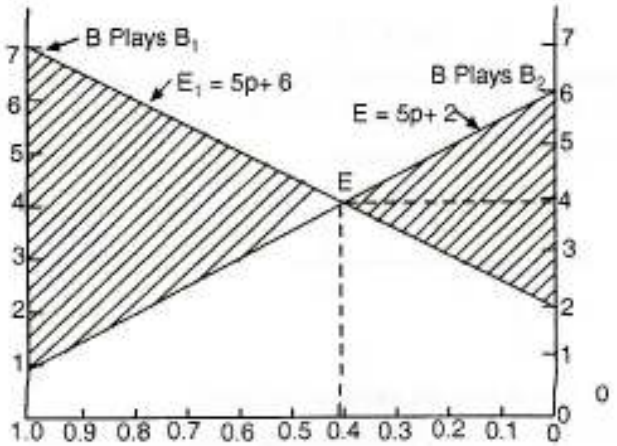
Assume that player B plays strategy B_1 all the time. What will the value of such a game be to player A? The value will depend on what player A does. If player A plays strategy A_1 all the time, the value of the game will be 1. If player A plays strategy A_2 all the time, the value of game will be 6 ; and if he mixes his strategies with a probability p and $(1-p)$, then the expected value of the game will be given by :

$$E_1 = p + 6(1-p) = -5p + 6$$

Similarly, if player B plays strategy B_2 all the time, value of the game to player A will be given by

$$E_2 = 7p + 2(1-p) = 5p + 2$$

The value $E_1 = -5p + 6$ and $E_2 = 5p + 2$ can be represented graphically by straight lines as shown in the figure given below :



The horizontal axis which represents the probability that player A will play strategy A_1 is constructed first, starting with $p = 0$ on the right side up to $p = 1$ on the left side. Every point on the horizontal axis represents the probability mix between strategy A_1 with probability 0.6 and choosing strategy A_2 with probability 0.4.

The vertical axis represents the payoff to player A. Now, two straight lines are plotted, each corresponding to one of player B's alternatives in the following way. For the payoff resulting from player B playing strategy B_1 all the time, a straight line is drawn from 7 on the left side to 2 on the right side. This is the line $E_1 = p + 6(1-p) = -5p + 6$. Next, a straight line is drawn for the situation in which player B plays strategy B_2 all the time; from 1 on the left to 6 on the right, describing the line $E_2 = 7p + 2(1-p) = 5p + 2$. The shaded area represents the entire feasible area of expected payoffs when p varies between 0 and 1. Player A will try to look for a payoff on the upper boundary of this feasible area. Player B will try to establish the payoff at the lower boundary of the feasible area. At point E, no matter what player B does, the payoff to player A is the maximum. However, it is also equal to the minimum payoff, *i.e.*, point E is the lowest on the upper boundary and the highest on the lower boundary.

The probability p can be read off the horizontal axis as 0.4, *i.e.*, player A should play strategy A_1 40% of the time, and therefore, strategy A_2 60% of the time. The value of the game is found at point E and can be read off from the vertical axis as $V = 4$.

In a similar manner, it is possible to present player B's situation on an additional graph. It will reveal that player B's strategy is to mix strategy B_1, 50% of the time and strategy B_2, 50% of the time.

Dominance Principle

When any strategy is better than another for all cases, it is said to be dominating the other and it is clear that we can discard the latter one from our consideration. Dominated rows or columns are deleted to reduce the size of the matrix. For example,

		Player B		
		1	2	3
Player A	1	1	7	2
	2	6	2	7
	3	5	1	6

Here, it is clear that each of A's payoffs for strategy 3 is less than that of strategy 2. Thus, strategy 3 is inferior and clearly strategy 2 is a dominating strategy. It appears that we might as well solve the game :

		Player B		
		1	2	3
Player A	1	1	7	2
	2	6	2	7

Also, if we observe, we find B's losses (A's payoffs) are always more for strategy 3 than for strategy 1, so B might as well never play strategy 3. Thus, our game reduces to

		Player B	
		1	2
Player A	1	1	7
	2	6	2

and thus the solution to our original game is the same as the solution to this game (so, we need only to solve for a 2×2 matrix).

Rule of dominance

(*a*) If all elements in a column are greater than or equal to the corresponding elements in another column, then that column is dominated.

(*b*) Similarly, if all elements in a row are less than or equal to the corresponding elements in another row, then that row is dominated.

MISCELLANEOUS ILLUSTRATIONS

Illustration 7. A proprietor of a food-stall has invented a new item of food delicacy which he calls WHIM. He has calculated that the cost of manufacture is Re. 1 per piece and that because of its novelty and quality it would be sold for Rs. 3 per piece. It is, however, perishable, and any goods unsold, at the end of the day are a dead loss. He expects the demand to be variable and has drawn up the following probability distribution expressing his estimates :

No. of pieces demanded :	10	11	12	13	14	15
Probability :	0.07	0.10	0.23	0.38	0.12	0.10

(*i*) Find an expression for his net profit or loss if he manufactures m pieces and only n are demanded. Consider separately the two cases $n \leq m$, and $n > m$.

(*ii*) Assume that he manufactures 12 pieces. Using the results in (*i*) above, find his net profit or loss for each level of demand.

(*iii*) Using the probability distribution, calculate his expected net profit or loss, if he manufactures 12 pieces.

(*iv*) Calculate similarly the expected profit or loss for each of the other levels of manufactures ($10 \leq m \leq 15$).

(*v*) How many pieces should be manufactured so that his net expected profit is maximum ?

Solution. (*i*) The proprietor does not produce more than 15 pieces of WHIM or less than 10 pieces. His profit is determined by the demand (n) and production (m). When the demand is more than the production, his profit shall be

$$\text{Rs. } 3 \times m - \text{Rs. } 1 \times m = \text{Rs. } 2m \qquad (\text{if } n > m)$$

When the production equals or exceeds the demand, his profit shall be

$$\text{Rs. } 3 \times n - \text{Rs. } 1 \times m = [\text{Rs. } 3n - m] \qquad (\text{if } n \leq m)$$

PAYOFF TABLE

Production m / *Demand* n	S_1 *10*	S_2 *11*	S_3 *12*	S_4 *13*	S_5 *14*	S_6 *15*	*Probability*
10	Rs. 20	Rs. 19	Rs. 18	Rs. 17	Rs. 16	Rs. 15	0.7
11	20	22	21	20	19	18	0.10
12	20	22	24	23	22	21	0.23
13	20	22	24	26	25	24	0.38
14	20	22	24	26	28	27	0.12
15	20	22	24	26	28	30	0.10

(*ii*) The S_j column of the payoff table given above, shows the net profit for each level of demand.

(*iii*) If he manufactures 12 pieces, his expected profits will be as follows :

$$0.07 \times 18 + 0.10 \times 21 + 0.23 \times 24 + 0.38 \times 24 + 0.12 \times 24 + 0.10 \times 24 = \text{Rs. } 23.28$$

(*iv*) The expected profits for other levels of manufacture are calculated below :

$$E(S_1) = [0.07 \times 20 + 0.10 \times 20 + 0.23 \times 20 + 0.38 \times 20 + 0.12 \times 20 + 0.10 \times 20] = \text{Rs. } 20$$

$$E(S_2) = [0.07 \times 19 + 0.10 \times 22 + 0.23 \times 22 + 0.38 \times 22 + 0.12 \times 22 + 0.10 \times 22] = \text{Rs. } 21.79$$

$$E(S_6) = \text{Rs. } 23.28 \text{ [Calculated above]}$$

$$E(S_4) = [0.07 \times 17 + 0.10 \times 20 + 0.23 \times 23 + 0.38 \times 26 + 0.12 \times 26 + 0.10 \times 26] = \text{Rs. } 24.08$$

$$E(S_5) = [0.07 \times 16 + 0.10 \times 19 + 0.23 \times 22 + 0.38 \times 25 + 0.12 \times 28 + 0.10 \times 28] = \text{Rs. } 23.74$$

$$E(S_6) = [0.07 \times 15 + 0.10 \times 18 + 0.23 \times 21 + 0.38 \times 24 + 0.12 \times 27 + 0.10 \times 30] = \text{Rs. } 23.04$$

(*v*) From the above calculations it is clear that he should manufacture 13 pieces for maximising his expected profit which is Rs. 24.08.

Illustration 8. Under an employment promotion programme it is proposed to allow sale of newspapers on the buses during off peak hours. The vendor can purchase the newspapers at a special concessional rate of Rs. 1.25 per copy against the selling price of Rs. 1.40. Any unsold copies are, however, a dead loss. A vendor has estimated the following probability distribution for the number of copies demanded :

Number of copies :	15	16	17	18	19	20
Probability :	0.04	0.19	0.33	0.26	0.11	0.07

How many copies should the vendor buy for maximum gain?

Solution. The vendor does not purchase less than 25 copies or more than 20 copies. His profit is determined by the demand (D) and the number of copies purchased (P). When the demand is more than the number of copies purchased by his profit will be :

$$\text{Rs. } 1.40 \times P - \text{Rs. } 1.25 \times P = \text{Rs. } 0.15\ P \quad \ldots(i)$$

When the demand is less or equal to the number of copies purchased by him the profit is :

$$\text{Rs. } 1.40 \times D - \text{Rs. } 1.25 \times P \quad \ldots(ii)$$

From (i) and (ii) we can make the following payoff table :

PAYOFF TABLE

No. of copies purchased P / *Demand D*	S_1 15	S_2 16	S_3 17	S_4 18	S_5 19	S_6 20	Probability
15	Rs. 2.25	Rs. 2.00	Rs. 1.75	Rs. 1.50	Rs. 1.85	Rs. 1.00	0.04
16	2.25	2.40	2.15	1.90	1.65	1.40	0.19
17	2.25	2.40	2.55	2.30	2.05	1.80	0.33
18	2.25	2.40	2.55	2.70	2.45	2.20	0.26
19	2.25	2.40	2.55	2.70	2.85	3.00	0.11
20	2.25	2.40	2.55	2.70	1.85	3.00	0.07

His expected profits are calculated below :

$E(S_1) = [2.25 \times 0.04 + 2.25 \times 0.19 + 2.25 \times 0.33 + 2.25 \times 0.26 + 2.25 \times 0.11 + 2.25 \times 0.07] = \text{Rs. } 2.25$

$E(S_2) = [2.00 \times 0.04 + 2.40 \times 0.19 + 2.40 \times 0.33 + 2.40 \times 0.26 + 2.40 \times 0.11 + 2.40 \times 0.07] = \text{Rs. } 2.38$

$E(S_3) = [1.75 \times 0.04 + 2.15 \times 0.19 + 2.55 \times 0.33 + 2.55 \times 0.26 + 2.55 \times 0.11 + 2.55 \times 0.07] = \text{Rs. } 2.44$

$E(S_4) = [1.50 \times 0.04 + 1.90 \times 0.19 + 2.3 \times 0.33 + 2.70 \times 0.26 + 2.70 \times 0.11 + 2.70 \times 0.07] = \text{Rs. } 2.37$

$E(S_5) = [1.85 \times 0.04 + 1.65 \times 0.19 + 2.05 \times 0.33 + 2.45 \times 0.26 + 2.85 \times 0.11 + 1.85 \times 0.07] = \text{Rs. } 2.144$

$E(S_6) = [1.00 \times 0.04 + 1.40 \times 0.19 + 1.80 \times 0.33 + 2.20 \times 0.26 + 3.00 \times 0.11 + 3.00 \times 0.07] = \text{Rs. } 2.012$

It is clear from the above calculations that the expected profit is maximum in the third course of action. Hence he should order 17 copies in order to maximise his profit.

Illustration 9. The probability of the demand for trucks for hiring on any day in a given district is as follows :

No. of trucks demanded	:	0	1	2	3	4
Probability	:	0.1	0.2	0.3	0.2	0.2

Trucks have a fixed cost of Rs. 90 each day to keep and daily hire charge (net or variable cost of running) is Rs. 200. If the Trucks-hire company owns 4 Trucks, what is its daily expectation? If the company is about to go into business and currently has no Trucks how many Trucks should it keep ?

Solution. Since the fixed cost of keeping each Truck is Rs. 90 a day, whether it is demanded or not and since, when demanded, it fetches a hire charge (net or variable cost of running) of Rs. 200, the payoffs with 4 Trucks are as under :

No. of Trucks demanded	:	0	1	2	3	4
Payoff (with 4 Trucks)	:	$0 - 90 \times 4 = -360$	$200 - 90 \times 4 = -160$	$400 - 90 \times 4 = 40$	$600 - 90 \times 4 = 240$	$800 - 90 \times 4 = 440$

Now the daily expectation is obtained by taking the sum total of the product of the above payoff and the corresponding probabilities of demand, *i.e.*,

$(-360)(0.1) + (-160)(0.2) + (40)(0.3) + (240)(0.2) + (440)(0.2) = \text{Rs. } 80$

For answering the other part of the question the payoff table is developed as under :

PAYOFF TABLE

Event (demand)	*Probability*	*Expected payoff (Rs.) Decision to Purchase Trucks (courses of action)*				
		0	1	2	3	4
0	0.1	0	–90	–100	–270	–360
1	0.2	0	+110	+20	–70	–160
2	0.3	0	+110	+220	+130	+40
3	0.2	0	+110	+220	+330	+240
4	0.2	0	+110	+210	+330	+440

COMPUTATION OF EMV OF VARIOUS COURSES OF ACTION

Event (demand)	*Probability*	*Expected payoff (Rs.) Decision to Purchase lorries (courses of action)*				
		0	1	2	3	4
0	0.8	0	–9	–18	–27	–36
1	0.2	0	22	4	–14	–32
2	0.3	0	33	66	39	12
3	0.2	0	22	44	66	48
4	0.2	0	22	44	66	88
EMV		0	90	140	130	80

Since EMV for the second course of action is highest (Rs. 140), therefore, the company should buy 2 lorries.

Illustration 10. A retailer purchases cherries every morning at Rs. 50 a case and sells for Rs. 80 a case remaining unsold at the end of the day can be disposed of next day at a salvage value of Rs. 20 per case (thereafter they have no value). Past sales have ranged from 15 to 18 cases per day. The following is the record of sales for past 120 days :

Cases	:	15	16	17	18
No. of days	:	12	24	48	36

Find how many cases the retailer should purchase per day to maximise his profit.

Solution. Here, number of cases of cherries purchased is an act or course of action and daily demand of the cherries is an event or state of nature. Using the information of the data, the various conditional profit (payoff) values for each act-event combination are given by :

Conditional (payoff) profit = (marginal profit) (cases sold) – (marginal loss) (cases unsold)

= (80–50) (cases sold) – (50–20) (cases unsold)

The resulting conditional payoff and corresponding expected payoff are computed in the following table :

Event (demand) per week	*Probability*	*Conditional payoff (Rs.) Act (Purchases per week)*				Expected payoff (*Rs.*) *Act (Purchases per week)*			
	(1)	15 (2)	16 (3)	17 (4)	18 (5)	15 (1) × (2)	16 (1) × (3)	17 (1) × (4)	18 (1) × (5)
15	0.1	450	420	390	360	45	42	39	36
16	0.2	450	480	450	420	90	96	90	84
17	0.4	450	480	510	480	180	192	204	192
18	0.3	450	480	510	540	135	144	153	162
	Expected Monetary Value (EMV) :					450	474	486	474

Since the act 'purchase 17 cases' yields the highest EMV of Rs. 486, the optimal act for the retailer would be to purchase 17 cases of cherries every morning.

Illustration 11. In a small town, there are two discount stores *ABC* and *XYZ*. They are the only stores that handle sundry goods. The total number of customers is equally divided between the two, because the price and quality of goods sold are equal. Both stores have good reputation in the community, and they render equally good customer services. Assume that a gain of customers by *ABC* is a loss to *XYZ*, and *vice versa*. Both stores plan to run annual pre-Diwali sales during the first week of October. Sales are advertised through the local newspaper, radio and television given below. (Figures in the matrix represent a gain or loss of customers). Find the optimal strategies for both stores and the value of game.

		Strategies of *XYZ*		
		Newspaper	Radio	Television
Strategies of *ABC*	Newspaper	30	40	–80
	Radio	0	15	–20
	Television	90	20	50

(MBA, DU, Nov. 2001; MBA, DU, Oct. 2003)

Solution. We observe that every element in the second row is less than the corresponding element in the third row. Applying the dominance principle, we shall delete the second row. Similarly, every element in the first column is greater than the corresponding element in the third column. Therefore, we delete the first column. Thus, the matrix is reduced to

		Strategies of *XYZ*	
		Radio	Television
Strategies of *ABC*	Newspaper	40	–80
	Television	20	50

Since there is no saddle point in this pay-off matrix, both stores will adopt a mixed strategy method. Let x be the probability of *ABC* for adopting Newspaper strategy. Therefore, probability of adopting Television strategy is $(1-x)$. Similarly, for store *XYZ*, let y and $(1-y)$ be the respective probabilities of adopting Radio and Television strategies. Thus,

$$E = 40xy - 80x(1-y) + 20(1-x)y + 50(1-x)(1-y)$$

Differentiating partially with respect to x and equating to zero, we get

$$\frac{\partial E}{\partial x} = 40y - 80(1-y) - 20y - 50(1-x)(1-y) = 0$$

or, $$y = \frac{13}{15} \quad \text{and } 1-y = \frac{2}{15}$$

Now, differentiating E partially with respect to y, we get

$$\frac{\partial E}{\partial y} = 40x + 80x + 20(1-x) - 50(1-x) = 0$$

or, $$x = \frac{1}{5} \quad \text{and } 1 - x = \frac{4}{5}$$

Substituting these values in the expression E, we get the value of the game, *i.e.*,

$$E = 40 \times \frac{1}{5} \times \frac{13}{15} - 80 \times \frac{1}{5} \times \frac{2}{15} + 20 \times \frac{4}{5} \times \frac{13}{15} + 50 \times \frac{4}{5} \times \frac{2}{15} = 24$$

Hence, the optimal strategies for both stores are that store *ABC* should go for Newspaper 20% times, no radio advertisement and Television should be adopted 80% times. Similarly, store *XYZ* should not go for Newspaper, 87% times Radio and 13% times Television. The value of the game will be 34 in favour of store *ABC*.

Illustration 12. Two breakfast food manufacturers *ABC* and *XYZ* are comparing for an increased market share. The payoff matrix shown in the following table, shows the increase in the share for *ABC* and decrease in market share of *XYZ* :

		XYZ			
		Give Coupons	Decrease price	Maintain present strategy	Increase advertising
ABC	Give coupons	2	–2	4	1
	Decrease price	6	1	12	3
	Maintain present strategy	–3	2	0	6
	Increase advertising	2	–3	7	1

Simplify the problem by the rule of dominance and then find optimal strategies for both the manufacturers and the value of the game.

Solution. Applying the rule of dominance, the payoff matrix can be reduced to

		XYZ	
		Give Coupons	Decrease Price
ABC	Decrease Price	6	1
	Maintain present strategy	–3	2

Since the payoff matrix cannot be further reduced, therefore, there is no saddle point. In this problem, both manufacturers will mix up the strategies in a random fashion. Let x and $(1-x)$ be the probabilities that the manufacturer *ABC* adopts in using strategies, decrease price and maintain present strategy respectively. Similarly, for manufacturer *XYZ*, let y and $(1-y)$ be the probabilities for using strategies, give coupons and decrease price respectively. Therefore, the expectation of manufacturer *ABC* is given by :

$$E = 6xy + x(1-y) - 3(1-x)y + 2(1-x)(1-y)$$

For maximizing the expectation, the first partial derivative must be equal to zero.

$$\frac{\partial E}{\partial x} = 6y + (1-y) - 3y - 2(1-y) = 0$$

or, $$y = \frac{1}{10}; \text{ and } (1-y) = \frac{9}{10}$$

$$\frac{\partial E}{\partial y} = 6x - x - 3(1-x) - 2(1-x) = 0$$

or, $$x = \frac{1}{2}; \text{ and } (1-x) = \frac{1}{2}$$

To obtain the value of the game, substitute the values of x and y in the expression E.

$$E = 6 \times \frac{1}{2} \times \frac{1}{10} + \frac{1}{2} \times \frac{9}{10} - 3 \times \frac{1}{2} \times \frac{1}{10} + 2 \times \frac{1}{2} \times \frac{9}{10} = 1.5$$

Hence the optimal strategies for both the manufacturers are that manufacturer *ABC* should adopt strategy 'decrease price' 50% times and strategy 'maintain present strategy, 50% times. Similarly, manufacturer *XYZ* should adopt strategy 'give coupons' 10% times and strategy 'decrease price' 90% times. The value of the game would be in favour of manufacturer *ABC* and the increase in markets share would be 1.5.

PROBLEMS

1-A: Answer the following questions, each question carries **one** mark:

(*i*) What is Statistical decision theory ?
(*ii*) What is payoff table ?
(*iii*) What is opportunity loss table ?
(*iv*) What is EVPI ?
(*v*) What is the difference between course of action and state of nature ?
(*vi*) What is decision-making under risk ?
(*vii*) What is decision-making under uncertainty ?
(*viii*) What is the difference between pure strategy and mixed strategy ?
(*ix*) What do you understand by dominance principle ?
(*x*) What is graphical method of a two-person zero-sum game ?

1-B: Answer the following questions, each question carries **four** marks:

(*i*) Explain briefly, the ingredients of a decision problem with suitable examples.
(*ii*) Describe any two methods of decision-making under uncertainty, pointing out their relative merits and demerits.
(*iii*) What is a two-person zero-sum game ? What are its major limitations ?
(*iv*) Differentiate between maximin and minimax principle.
(*v*) List the different steps in decision-making. (*MBA, Madras Univ., Nov. 2003*)
(*vi*) Explain the difference between decision-making under certainty, risk and uncertainty by giving suitable example.
(*vii*) Describe at least two methods of solving a two-person zero-sum game problem.

2. Explain how statistics is useful in the decision-making process of business and management.

3. How is Expected Value calculated? What are the advantages and disadvantages of using Expected Value as a decision criterion ?

4. (*a*) Decision criteria under situation of uncertainty is governed by the attitude of the decision-maker." Explain.
(*b*) Describe some methods which are useful for decision-making under uncertainty. Illustrate each by an example.

5. Explain clearly the following :
(*i*) Course of action (*ii*) State of nature
(*iii*) Payoff table (*iv*) Opportunity loss. (*MBA, HPU, 2002*)

6. Explain the following, giving a suitable example :
(*i*) The minimax principle (*ii*) The maximin principle, (*iii*) The Bayes principle (*iv*) Expected value of perfect information. (*v*) Highest Expected payoffs with information, (*vi*) Highest expected payoffs under uncertainty.

7. Explain the difference between expected opportunity loss and expected value of perfect information.

8. Explain the maximin and minimax regret criteria of decision-making under uncertainty giving suitable examples.

9. What is meant by 'Statistical Decision Theory'. How is it different from other methods used in decision-making? Describe some methods which are useful for decision-making under uncertainty.

10. Explain the following terms :
    (*i*) Two-person zero-sum games,
    (*ii*) Principle of dominance, and
    (*iii*) Pure strategy in game theory.
11. What is game theory? Include in your answer various approaches in solving for strategies and game values.
12. What is two-person zero-sum game? What are its major limitations ?
13. Explain the terms : minimax strategies, saddle point, mixed strategies and principle of dominance.
14. "The primary contribution of the game theory has been its concept rather than its formal application to solving real-life problems." Do you agree? Discuss.
15. (*a*) "Game theory deals with making decisions under conflict caused by opposing interests." Elucidate this statements by giving appropriate examples.
    (*b*) Explain the criterion of maximin and minimax regret in the context of decision theory. *(M. Com., DU, 1999)*
16. A baker makes a certain kind of pastry at night and sells it the next day. It is perishable and must be thrown if not sold during the day. The unit cost and price of the pastry are Re.1 and Rs. 3 respectively. According to the past experience, the daily demand (in hundred) and the respective probabilities are :

Demand	:	20	21	22	23	24
Probability	:	0.1	0.2	0.3	0.3	0.1

    (*i*) Construct the payoff table.
    (*ii*) Construct the loss table.
    (*iii*) Determine the maximin and maximax action.
    (*iv*) Compute the highest expected payoff with perfect information.
17. A certain product is manufactured at Rs. 50 and sold at Rs. 75 per unit. The product is such that if it is produced but not sold during a weeks' time, it becomes worthless. The weekly sales records in the past are as follows :

Demand per week	:	20	21	22	23
No. of weeks each sales level was recorded	:	200	350	800	150

    (*i*) Calculate the expected sales of the month.
    (*ii*) Prepare a table of payoff for different possible acts.
    (*iii*) Prepare a table of expected payoffs and select the optimal act.
18. A stall agent at a certain railway station sells for Rs. 1.50 a copy of daily newspaper for which it repays Rs. 1.22. Unsold papers are returned for a refund of 50 paise a copy. The daily sales and corresponding probabilities are as follows :

Daily sales	:	500	600	700
Probability	:	0.5	0.3	0.2

    (*i*) How many copies should he order each day ?
    (*ii*) If unsold copies cannot be returned and are useless, what should be optimal order each day?
19. N. Sombhai & Co., a wholesale dealer in electrical appliances, was offered an agency for selling Godrej refrigerators. The company estimated that his fixed costs in taking up the agency would be Rs. 1,20,000 per year. Contribution per refrigerator sold would be Rs. 2,000. From a potential target audience of 2,000 buyers in the region, the company assessed that its market for the refrigerator sales would be 2%, 4% or 6% of the target audience with probability 0.1, 0.6 and 0.3 respectively. For a fixed cost of Rs. 10,000, the company could get a sample survey of potential buyers conducted. Whether the company should go in for additional information? Support your decision with appropriate reason in terms of EMV.
20. A company is currently involved in negotiations with its union on the upcoming wage contract. With the aid of an outside mediator, the table below was constructed by the management group. The plus points are to be interpreted as proposed wage increases while a minus figure indicates that a wage reduction is proposed. The mediator informs the management group that he has been in touch with the union and they have constructed a table that is comparable to the table developed by management. Both the company and the union must decide on an overall strategy, before negotiation. The management group understands the relationships of company strategies to union strategies in the following table but lacks specific knowledge of game theory to select the best strategy (or strategies) for the company. You have been called in to assist management on this problem. What game value and strategies are available to the opposing groups?

CONDITIONAL COST (Rs.) TO COMPANY

Union Strategies

		U_1	U_2	U_3	U_4
Company Strategies	C_1	+0.25	+0.25	+0.35	– 0.02
	C_2	+0.20	+0.16	+0.08	+0.08
	C_3	+0.14	+0.12	+0.15	+0.13
	C_4	+0.30	+0.14	+0.19	0

21. *A* and *B* play a game in which each has three coins, a 5p., a 10p. and a 20p. Each selects a coin without the knowledge of other's choice. If the sum of the coins is an odd amount, *A* wins *B*'s coin; if the sum is even, *B* wins *A*'s coin. Find the best strategy for each player and the value of the game. *(MBA, Delhi Univ., 1997)*

22. A businessman has three alternatives open to him and each of which can be allowed by any of the four possible events. The conditional payoffs for each action event combination are given below :

	Event			
Action	*A*	*B*	*C*	*D*
S_1	8	0	–10	6
S_2	– 4	12	18	–2
S_3	14	6	0	8

(*a*) If he adopts maximin criterion, what act he should choose?

(*b*) If the criterion of choice is minimax regret; what action should be chosen ?

(*c*) If he uses EMV (Expected Monetary Value) as his decision criterion, what action should he choose (assume that events have equal probability of occurrence)?

23. A person has the choice of running a hot snack stall or an ice-cream and cold drink shop at a certain holiday resort during the coming summer season. If the weather during the season is cool and rainy he can expect to make a profit of Rs. 15,000 and if it is warm he can expect to make a profit of only 3,000 by running a hot snack stall. On the other hand, if his choice is to run an ice-cream and cold drink shop he can expect to make a profit of Rs. 18,000, if the weather is warm and only Rs. 3,000 if the weather is cool and rainy. The meteorological authorities predict that there is 40% chance of the weather being warm during the coming season. You are to advice him as to the choice between the two types of stalls. Base clearly your argument on the expectation of the results of the two courses of action and show the result in a tabular form.

24. Vishal who possesses an amount of Rs. 1 lakh is planning to invest it among three companies : equity shares in company A,B and C. The payoff terms of (i) growth in capital and (ii) returns to capital are known for each of the investments under each of the three economic conditions which may prevail, that is, recession, growth and stability. Assuming that Vishal must make his choice among the three portfolios for a period of one year in advance, his expectations of the net earning (in Rs '000) of his Rs. 1 lakh portfolio after one year is represented by the following matrix :

	Recession	*Stability*	*Growth*
Company *A*	–15	6	10
Company *B*	4	7.5	8
Company *C*	6.5	6	5

Determine the optimal strategies for investment and the expected per cent return for the investor under such a policy.

25. A physician purchases a particular vaccine on Monday each week. The vaccine must be used within the following quantities, otherwise it becomes worthless. The vaccine costs Rs. 5 per dose and the physician charges Rs. 10 per dose. In the past 50 weeks, the physician has administered the vaccine in the following quantities :

Dose per week	:	20	25	50	60
Number of weeks	:	5	15	25	5

Determine how many doses the physician should buy every week.

26. Given is the following payoff matrix :

State of nature	*Probability*	*(Decision Rs.)* *Do not expand*	*Expand*	*Expand*
		100 units	*200 units*	*400 units*
High demand	0.4	2500	3500	5000
Medium demand	0.4	2500	3500	1000
Low demand	0.2	2500	1500	1000

What should be the decision, if we use (*i*) *EMV* criterion (*ii*) the maximin criterion, (*iii*) the maximax criterion and (*iv*) minimax regret criterion ?

27. A toy camera manufacturer produces two models (standard and deluxe). In preparation for the heavy Christmas selling season, he must decide how many of each model to produce. Variable cost of standard camera is Rs. 10 and selling price is Rs. 20; variable cost of the deluxe model is Rs. 35. He estimates demand as follows :

Standard model		*Deluxe model*	
Demand	*Probability*	*Demand*	*Probability*
6,000	0.30	2,000	0.20
8,000	0.70	4,000	0.80

Any camera not sold during the season is sold at salvage price of Rs. 5 for the standard camera and Rs. 10 for the deluxe camera. The manufacturer feels that different segments of the market purchase the two different models, thus the probabilities of sales given above are independent. Assuming unlimited production capacity, the two decisions can be made independently. What are the optimal quantities of each model to produce? What are the two optimal EMV's? (*MBA, M.D. Univ., 1997*)

28. Crown Auto is trying to decide about the size of the plant to be built in Noida. Three alternatives of annual capacity, *viz.*, (*i*) 10,000 units (*ii*) 20,000 units and (*iii*) 30,000 units are under consideration. Demand for the product is not known with certainty but the management has estimated the probabilities for 5 different levels of demand. The profit for each size of plant at different levels of demand is as follows :

Level of demand	*Probability*	*Decision (Rs. in lakhs)*		
		10,000 units	*20,000 units*	*30,000 units*
Very high	0.15	– 0	– 6	– 8
High	0.30	1	0	– 2
Moderate	0.25	1	7	5
Low	0.20	1	7	11
Very low	0.10	1	7	11

What plant capacity would you suggest to the management? Also find EVPI.

29. Two leading firms Nirmala Textiles Ltd., and Swati Rayons Ltd., for years have been selling, shirting which is but a small part of both firms total sales. The Marketing Director of Nirmala Textiles raised the question: "What should his firm's strategies be in terms of advertising for the product in question?" The system group of Nirmala Textiles developed the following data for varying degrees of advertising :

(*i*) No advertising, medium advertising and heavy advertising for both firms will result in equal market share.

(*ii*) Nirmala Textiles with no advertising : 40 per cent of the market with medium advertising by Swati Rayons and 28 per cent of the market with heavy advertising by Swati Rayons.

(*iii*) Nirmala Textiles using medium advertising : 70 per cent of the market with no advertising by Swati Rayons and 45 per cent of the market with heavy advertising by Swati Rayons.

(*iv*) Nirmala Textiles using heavy advertising ; 75 per cent of the market with no advertising by Swati Rayons and 52.5 per cent of market with medium advertising by Swati Rayons.

Based upon the above information, answer the marketing director's question.

30. A newspaper boys buys papers for Rs. 1.75 each and sells them for Rs. 1.95 each. He cannot return unsold newspapers. Daily demand has the following distribution :

No. of customers	:	220	221	222	223	224	225	226	227	228
Probability	:	0.22	0.03	0.05	0.05	0.25	0.05	0.20	0.10	0.05

If each day's demand is independent of the previous day's demand, how many newspapers should be ordered each day ?

31. Solve the following two-person zero-sum game :

		Player *B*				
		1	2	3	4	5
Player *A*	1	2	– 4	– 6	– 3	5
	2	– 3	4	– 4	1	0

32. Two firms are competing for business. Whatever firm A gains, firm B loses. The table below shows advertising strategies of both firms and the utilities to firm A for various market shares in percentages.

		Firm B		
		Press	Radio	TV
Firm A	Press	60	45	40
	Radio	75	75	60
	TV	80	60	70

Find optimal strategies for both firms and expected percentage of market shares of firm A.

33. A small industry finds from the past data that the cost of making an item is Rs. 25, the selling price of an item is Rs. 30, if it is sold within a week, and it could be disposed of at Rs. 20 per item at the end of the week.

Weekly Sales :	≤ 3	4	5	6	7	≥ 8
No. of Weeks :	0	10	20	40	30	0

Find the optimum number of items per week the industry should produce.

34. A management is faced with the problem of choosing one of three products for manufacturing. The potential demand for each product may turn out to be good, moderate or poor. The probabilities for each of states of nature were estimated as follows :

Product	*Nature of Demand*		
	Good	*Moderate*	*Poor*
X	0.70	0.20	0.10
Y	0.50	0.30	0.20
Z	0.40	0.50	0.10

The estimated profit or loss under the various types of nature of demand may be taken as :

	Rs.	*Rs.*	*Rs.*
X	30,000	20,000	10,000
Y	60,000	30,000	20,000
Z	40,000	10,000	–15,000

Prepare the expected monetary value table and advise the management about the choice of product.

35. Solve the following game by using the principle of dominance :

		Player B					
		I	II	III	IV	V	VI
Player A	1	4	2	0	2	1	1
	2	4	3	1	3	2	2
	3	4	3	7	–5	1	2
	4	4	3	4	–1	2	2
	5	5	3	3	–2	2	2

36. Solve the following two-person zero-sum game :

		Player B		
		B_1	B_2	B_3
Player A	A_1	4	5	8
	A_2	–2	–3	4
	A_3	–6	–4	0
	A_4	6	–5	2

37. A firm makes pastries, which it sells for Rs. 8 per piece in special boxes containing one dozen each. The direct cost of pastries for the firm is Rs. 4.50 per piece. At the end of the week, the stale pastries are sold off for a lesser price of Rs. 2.50 per piece. The overhead expense attributable to the pastry production is Rs. 1.25 per piece. Fresh pastries are sold in special boxes which cost 50 paise each and the stale pastries are sold wrapped in ordinary paper. The probability distribution of demand per week is as under :

Demand (in dozen)	:	0	1	2	3	4	5
Probability	:	0.01	0.14	0.20	0.50	0.10	0.05

Find the optimal production level of pastries per week.

38. Firm X is fighting for its life against the determination of firm Y to drive it out of the industry. Firm X has the choice of increasing its price, leaving it unchanged or lowering it. Firm Y has the same three options. Firm X's gross sales in the event of each of the possible pairs of choice are shown below :

		Firm Y's Pricing Strategy		
		Increase price	*Do not change*	*Reduce price*
Firm X's Pricing Strategy	Increase Price	90	0	110
	Do not change	110	100	90
	Reduce price	120	70	80

Find the optimal strategies for both the firms and also the value of the game.

39. Assume that two firms are competing for market share for a particular product. Each firm is considering what promotional strategy to employ for the coming period. Assume that the following payoff matrix describes the increase in market share for firm A and the decrease in market share for firm B :

		Firm B		
		No Promotion	*Moderate Promotion*	*Heavy Promotion*
Firm A	No. Promotion	3	0	−3
	Moderate Promotion	2	3	1
	Heavy Promotion	−4	2	−1

Determine the optimal strategies for each firm and the value of the game.

40. For the following payoff matrix, find the value of the game and the strategies of players A and B by using graphical method :

		Player B		
		B_1	B_2	B_3
Player A	A_1	3	−1	4
	A_2	6	7	−2

41. The management of a corporation is in the process of deciding whether to agree to negotiate with the striking union, now or to delay. The decision is difficult because the management does not know the union leadership's position. The union leaders may be adamant and insist on their original demands, they may be ready to compromise or they may be ready to yield and accept the original management offer. The matrix of payoffs to management, as management sees it, is (in Rs. 1 million units) given below :

	UNION POSITION		
	B_1 *Adamant*	B_2 *Compromise*	B_3 *Yield*
A_1 *Negotiate*	−2	−1	2
A_2 *Delay*	5	−2	−3

(*i*) Solve management's problem.
(*ii*) What should be the union's strategy.
(*iii*) Discuss the implications of a conclusion to adopt a random strategy.

42. A producer of boats has estimated the following distribution of demand for a particular kind of boats :

No. demanded :	0	1	2	3	4	5	6
Probability :	0.14	0.27	0.27	0.18	0.09	0.04	0.01

Each boat cost him Rs. 70,000 and he sells for Rs. 1,00,000 each. Any left unsold at the end of the season must be disposed for Rs. 60,000 each. How many should he stock so as to maximise his expected payoff ?

43. There are two companies *A* and *B* in a certain city. Both companies have similar reputation and the total number of customers is equally divided between the two companies. Both the companies want to attract more number of customers by using different media of advertisement. By seeing the market trend, the company *A* constructed the following payoff matrix, where the numbers in the matrix indicate a gain or a loss of customers.

		Company B		
		Newspaper	*Radio*	*T.V.*
Company A	*Newspaper*	40	50	–70
	Radio	10	25	–10
	T.V.	100	30	60

Find optimal strategies for both the companies and value of the game.

44. A group of students raise money each year by selling Souvenirs outside the stadium after a cricket match between Teams *A* and *B*. They can buy any of the three different types of Souvenirs from a supplier. Their sales are mostly dependent on which team wins the match. A conditional payoff table is as under :

	Type of Souvenir		
	I	*II*	*III*
Team *A* wins	Rs. 1,200	Rs. 800	Rs. 300
Team *B* wins	Rs. 250	Rs. 700	Rs. 1,100

(*i*) Construct the opportunity loss table.

(*ii*) Which type of Souvenir should the students buy if probability of Team *A*'s winning is 0.6?

45. The conditional payoffs in rupees for each action-event combination are as under :

	Action			
Event	*1*	*2*	*3*	*4*
A	4	–2	7	8
B	0	6	3	5
C	–5	9	2	–3
D	3	1	4	5
E	6	6	3	2

(*i*) Which is the best action in accordance with the maximin criterion?

(*ii*) Which is the best action in accordance with EMV criterion, presuming all events have equal probabilities of occurrence?

46. In a duopolistic market, two competitor compete for profit with promotional effort as their only controllable variables. Each competitor has the option of increasing or decreasing the promotional expenditure or staying at the normal level. The expected increase in profit of competitor 1 under various situations is shown here (in Rs. 10,000 units) :

		Competitor 2		
		Increase	*Normal*	*Decrease*
Competitor 1	*Increase*	–200	–20	30
	Decrease	–50	20	40
	Normal	80	10	50

Assuming a zero-sum game, find the optimal strategy of each competitor and the value of the game.

47. Two companies *A* and *B* are competing for the similar type of product. Their different strategies are given in the following payoff matrix.

		Company *B*		
		B_1	B_2	B_3
Company A	A_1	2	–2	3
	A_2	–3	5	–1

Determine the best strategies for both the companies and also the value of the game.

48. In a recreation beach, two persons, A and B, are interested in starting a refreshment stall. Initially, only three places are under consideration. The following payoff matrix for different strategies of the players is given :

		B's position		
		Entrance	*Centre*	*Exit*
A's Position	*Entrance*	50	30	40
	Centre	70	50	60
	Exit	60	70	50

What is the best strategy for A and B to start the refreshment stall ?

49. A soft drink company calculated the market share of two products against its major competitor having three products and found out the impact of additional advertisement in any one of its product against the other.

		Competitor		
		1	2	3
Company	1	6	7	15
	2	20	12	10

What is the best strategy for the company as well as the competitor ? What is the payoff obtained by the company and the competitor in the long run? Use graphical method to obtain the solution.

50. Two candidates, X and Y, are competing for the councillors seat in a city municipal corporation, and X is attempting to increase his total votes at the expense of Y. The strategies available to each candidate involve personal contacts, newspapers insertions or television advertising. The increase in votes available to X given various combinations of strategies are given below. Assuming two-person zero-sum game, determine the optimal strategies that should be adopted by X during his election campaign. How many votes should X gain by the following optimal strategy ?

		Y		
		Personal Contacts	*Newspaper*	*Television*
X	*Personal contacts*	30,000	20,000	10,000
	Newspaper	60,000	50,000	25,000
	Television	20,000	40,000	30,000

51. A production manager has calculated that for every additional unit sold he makes an additional profit of Rs. 2, but for every unit left unsold, he loses Rs. 1.20. The probability distribution for the demand (in lakh units) of the product per week is given below :

Demand per week	:	20	21	22	23	24	25	26	27
Probability	:	0.24	0.08	0.09	0.17	0.15	0.13	0.09	0.05

Determine the optimal number of units the production manager should store for a week.

52. Assume that a manager sells an article having normally distributed sales with a mean of 50 units daily and a standard deviation in daily sales of 15 units.The manager purchases this article for Rs. 4 per unit and sells it for Rs.9 per unit. If the article is not sold on the selling day, it is worth nothing. Determine the optimal size of the order of the article, the manager should make daily. *(MFC, Delhi Univ., 1996)*

53. A big breeder can either produce 20 or 30 pigs. The total production of his competitors can be either 5,000 or 10,000 pigs. If they produce 5,000 pigs, his profit per pig is Rs. 60; if they produce 10,000 pigs, his profit per pig is only Rs. 45. Construct a payoff table and also state what would the big breeder decide.

54. Two firms A and B are competing for the same type of product. Their different strategies are given in the following payoff matrix :

		Firm B			
		B_1	B_2	B_3	B_4
Firm A	A_1	35	65	25	5
	A_2	30	20	15	0
	A_3	40	50	0	10
	A_4	55	60	10	50

Using the concept of dominance, reduce this game to 2×2 matrix. Also determine their optimal strategies and the value of the game.

55. For the following matrix, find the optimal strategies for A and B and the value of the game :

		Firm B		
		B_1	B_2	B_3
Firm A	A_1	12	10	8
	A_2	14	14	10
	A_3	16	12	15

56. Under an employment promotion scheme, it is proposed to allow sale of newspaper on the buses during off peak hours. A vendor can purchase the newspaper at a concessional rate of Rs. 1.70 per copy and sell it for Rs. 1.90. Copies unsold at the end of the day are, however, a dead loss. The demand probability distribution has been estimated as follows :

Demand	:	160	170	180	190	200	210
Probability	:	0.04	0.19	0.33	0.26	0.11	0.07

How many copies should the vendor order so as to maximise his expected profit ? *(M.Com., DU, 1999)*

57. Consider the following payoff (profit) matrix :

		State of nature				
		N_1	N_2	N_3	N_4	N_5
Strategy	S_1	60	70	–10	0	40
	S_2	30	45	20	35	–15
	S_3	40	35	25	20	30
	S_4	50	–20	35	25	20

No probabilities are known for the occurrence of the state of nature. Compare the solutions obtained by each of the following criteria: (*a*) Maximin, (*b*) Regret, (*c*) Laplace (*d*) Hurwicz. *(MBA, Madras Univ., 1999)*

58. Consider the following pay off (profit) matrix.

		State of Nature				
		N_1	N_2	N_3	N_4	N_5
Strategy	S_1	60	70	–10	0	40
	S_2	30	45	20	35	–15
	S_3	40	35	25	20	30
	S_4	50	–20	35	25	20

Compare the solutions obtained by Minimax (Savage) and Laplace criterion.

59. A company needs to increase its production beyond its existing capacity. It has narrowed the alternatives to two approaches to do so : (*a*) expansion at a cost of Rs. 8 million, or (*b*) modernization at a cost of Rs. 5 million. Both approaches would require the same amount of time for implementation. Management believes that over the required payback period, demand will either be high or moderate. Since high demand is considered to be somewhat less likely than moderate demand, the probability of high demand has been set at 0.35. If the demand is high, expansion would gross an estimated additional high demand has been set at 0.35. If the demand is high, expansion would gross an estimated additional Rs. 12 million but modernization only an additional Rs. 6 million, due to a lower maximum production capability. On the other hand, if the demand is moderate, the comparable figures would be Rs. 7 million for expansion and Rs. 5 million for modernization.

(*a*) Calculate the conditional profit in relation to variance action and outcome combinations and states of nature.
(*b*) If the company wishes to maximize its expected monetary value (*EMV*), should it modernize or expand ?
(*c*) Calculate the *EVPI*.
(*d*) Construct the conditional opportunity loss table and also calculate *EOL*.

60. Suppose an analysis of demand for a product in the last one year (52 weeks) revealed the demand distribution given in the table given below :

Table : Demand Distribution

Quantity demanded	*No. of weeks this quantity was sold*	*Probability*
30	5	0.10
31	10	0.20
32	16	0.30
33	13	0.25
34	5	0.10
35	3	0.05
	52	1.00

Selling price of the product = Rs. 3.00.

Cost price of the product = Rs. 2.00.

Selling price of more than one week old product = Rs. 1.00 (*i.e.*, loss of unsold unit)

(*i*) Construct the conditional profit table.

(*ii*) Determine the optimum number of units of his commodity, to order weekly in order to maximize his profit.

(*iii*) Compute *EPPI* and *EVPI*.

(*iv*) Construct the conditional loss table.

(*v*) Compute *EOL*.

(*vi*) Compare (*iii*) and (*v*).

61. A Company has to decide on marketing one of the following two types of portable transistor radios—Deluxe and Popular. The market forecast for the coming festival season indicates 75% chance that the market will be good, 15% chance it will be fair and 10% chance it will be poor. The payoffs for each strategy corresponding to the different states of nature is given in the following matrix :

States of Nature

	Market Good	*Market Fair*	*Market poor*
Probability	0.75	0.15	0.10
Strategy		*Pay offs (Rs.)*	
Deluxe Model	35,000	15,000	5,000
Popular Model	50,000	20,000	(–) 3,000

Which strategy the company should choose ?

62. Mr. Ram buys a perishable commodity at Rs. 5 each. The profit per unit is Rs. 5. This perishable commodity he ca keep in his shop for a week and at the end of each week the leftover are sold to a restaurant for Rs. 3 each (a loss of Rs. each). Mr. Ram has the record for past 100 weeks for his weekly sales as given below :

Weekly demand :	1	2	3	4	5	6	7
Number of weeks :	5	10	25	30	20	5	5

(*i*) Construct the conditional profit table.

(*ii*) Determine the optimum number of units of his commodity to order weekly in order to maximize his profit.

(*iii*) Compute *EPPI* and *EVPI*.

(*iv*) Construct the conditional loss table.

(*v*) Compute *EOL*.

(*vi*) Compare (*iii*) and (*v*).

*****

Madras University

MBA—Business Statistics

Time : Three hours Maximum : 75 marks

Part A *(10 × 1 = 10 marks)*

Answer any TEN questions. All questions carry equal marks.

1. What is probability ?
2. What is a decision making ?
3. Define the term research.
4. Define Sample.
5. What is hypothesis ?
6. Define editing.
7. What do you mean by cluster ?
8. What is Univariate ?
9. List out the types of revenues.
10. What is surplus ?
11. Define parametric.
12. What is nominal scale ?

Part B *(5 × 5 = 25 marks)*

Answer any FIVE questions. All questions carry equal marks.

13. State Baye's theorem and explain its applications.
14. What do you understand by decision tree ? Explain, how it will be useful in decision making ?
15. Discuss the scope and objectives of research.
16. List out and explain the different methods of data collection.
17. What do you mean by conjoint analysis ? Explain.
18. Explain different types of surpluses in detail.
19. State few applications of differentiation and integration.

Part C *(4 × 10 = 40 marks)*

Question no. 20 is compulsory. Answer any three questions from Q.21 to 24. All questions carry equal marks.

20. Five white and six red balls are in a bag. Two drawings of three balls are made such that
    (a) the balls are replaced before the second trial, and
    (b) the balls are not replaced before the second trial. Find the probability that the first drawing will give three white and the second drawing will give three red balls in each case.
21. In an experiment of pea-breeding Mendel obtained the following frequencies of seed : 315 round and yellow, 101 wrinkled and yellow, 108 round and green, 32 wrinkled and green. According to his theory of heredity, the numbers should be in proportion 9 : 3 : 3 : 1. Is there any evidence to doubt the theory at 5% level of significance? (Tabulated value : For $df = 3$, chi-square at 5% is 7.82)
22. Discuss different types of research used in management with examples.
23. The following data present the yields in quintals of common ten sub-divisions of equal area of two agricultural plots :

Plot 1 :	6.2	5.7	6.5	6.0	6.3	5.8	5.7	6.0	6.0	5.8
Plot 2 :	5.6	5.9	5.6	5.7	5.8	5.7	6.0	5.5	5.7	5.5

Test whether two samples taken from two random populations have the same variance.
(5% point of F for $v_1 = 9$ and $v_2 = 9$ is 3.18)

24. Give an account of report writing by explaining different types of reports

APPENDIX

Statistical Tables

I. Logarithms

II. Antilogarithms

III. Powers, Roots and Reciprocals

IV. Binomial Coefficients

V. Values of e^{-m}

VI. Ordinates (Y) of the Standard Normal Curve at Z

VII. Areas under the Standard Normal Distribution

VIII. Critical Values of χ^2

IX. Critical Values of t

X. 5% Points of F-distribution

XI. 1% Points of F-Distribution

XII. Control Charts Constants

XIII. Random Numbers

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	0128	0170	0212	0253	0294	0334	0374	5 4	9 8	13 12	17 16	21 20	26 24	30 28	34 32	38 36
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0775	4 4	8 7	12 11	16 15	20 18	23 22	27 26	31 29	35 33
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3 3	7 7	11 10	14 14	18 17	21 20	25 24	28 27	32 31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3 3	6 7	10 10	13 13	16 16	19 19	23 22	26 25	29 29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3 3	6 6	9 9	12 12	15 14	19 17	22 20	25 23	28 26
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3 3	6 6	9 8	11 11	14 14	17 17	20 19	23 22	26 25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3 3	6 5	8 8	11 10	14 13	16 16	19 18	22 21	24 23
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	3 3	5 5	8 8	10 10	13 12	15 15	18 17	20 20	23 22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2 2	5 4	7 7	9 9	12 11	14 14	17 16	19 18	21 21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2 2	4 4	7 6	9 8	11 11	13 13	16 15	18 17	20 19
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	3560	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	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5456	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	6064	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	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	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	6	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	5	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8

II. LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1 2 3	4 5 6	7 8 9
50	6990	6998	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	7226	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
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1 2 2	3 4 5	5 6 7
56	7582	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 5 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	7996	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	8269	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	8445	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 5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1 1 2	2 3 4	4 5 5
73	8633	8639	8645	8651	8557	8663	8669	8675	8681	8686	1 1 2	2 3 4	4 5 5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1 1 2	2 3 4	4 5 5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1 1 2	2 3 3	4 5 5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1 1 2	2 3 3	4 5 5
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	9031	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 5
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	9410	9415	9420	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	9595	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	9717	9722	9727	0 1 1	2 2 3	3 4 4
94	9731	9736	9741	9745	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	9859	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	9987	9991	9996	0 1 1	2 2 3	3 3 4

I. 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	1139	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
.12	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	0 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	1517	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 4	5 6 6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1 1 2	3 4 4	5 6 6

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	3350	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	2	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	4	5	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	4556	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	5164	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	5689	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	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	6353	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	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	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	6	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	8810	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

III. POWERS, ROOTS AND RECIPROCALS

n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$	$\sqrt{10n}$	$\sqrt[3]{10n}$	$\sqrt[3]{100n}$	$\frac{1}{n}$
1	1	1	1	1	3.162	2.154	4.642	1
2	4	8	1.414	1.260	4.472	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	7.638	.2500
5	25	125	2.236	1.710	7.701	3.684	7.937	.2000
6	36	216	2.449	1.817	7.746	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	256	4096	4.000	2.520	12.649	5.429	11.696	.06250
17	289	5913	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	2.668	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	1.04762
22	484	10648	4.690	2.802	14.832	6.037	13.006	.04545
23	529	12167	4.796	2.844	15.166	6.127	13.200	.04348
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	6.840	14.736	.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

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	3.826	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	8.434	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
66	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	357011	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	8.004	19.399	.01370
74	5476	405224	8.602	4.198	27.203	9.045	19.487	.01351
75	5625	421875	8.660	4.217	27.386	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	8.883	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	20.567	.01149
88	7744	681472	9.381	4.448	29.665	9.583	20.646	.01136
89	7921	704969	9.434	4.465	29.833	9.619	20.724	.01124
90	8100	729000	9.487	4.481	30.000	9.655	20.801	.01111
91	8281	753571	9.538	4.498	30.166	9.691	20.878	.01099
92	8464	778688	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	8836	830584	9.695	4.547	30.659	9.796	21.105	.01064
95	9025	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	9.909	4.626	31.464	9.967	21.472	.01010
100	10000	1000000	10.000	4.642	31.623	10.000	21.544	.0100

IV. BINOMIAL COEFFICIENTS

n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	3005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756

V. VALUES OF e^{-m} (For Computing Poisson Probabilities)

$(0 < m < 1)$

m	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	0.9048	.8958	.8860	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	0.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	0.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	0.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	0.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	0.5488	.5434	.5379	.5326	.5278	.5220	.5160	.5117	.5066	.5016
0.7	0.4966	.4916	.4868	.4810	.4771	.4724	.4670	.4630	.4584	.4538
0.8	0.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	0.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716

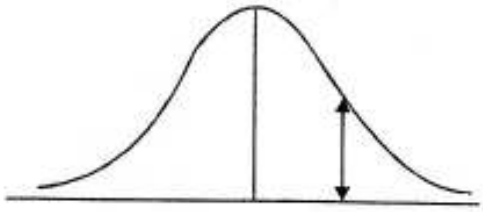
$(m = 1, 2, 3, \ldots\ldots, 10)$

m	1	2	3	4	5	6	7	8	9	10
e^{-m}	.36788	.13534	.04979	.01832	.006737	.002478	.00092	.000335	.000123	.000045

Note : To obtain values of e^{-m} for other values of m, use the laws of exponents.

Example. $e^{-2.35} = (e^{-2.00})\,(e^{-0.35}) = (.13534)\,(.7047) = .095374$

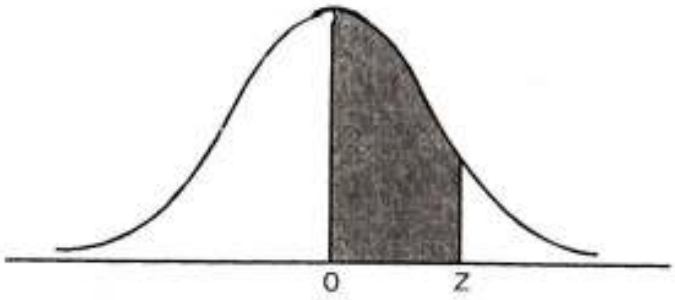
VI. ORDINATES (Y) OF THE STANDARD NORMAL CURVE AT z



z	0	1	2	3	4	5	6	7	8	9
0.0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3925	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3725	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2897	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1450	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0923	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2.0	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0194	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0051	0.0051	0.0050	0.0048	0.0047	0.0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006
3.6	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
3.7	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003
3.8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002
3.9	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001

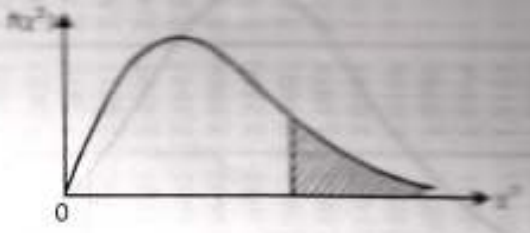
VII. AREAS UNDER THE STANDARD NORMAL DISTRIBUTION

The entries in this table are the probabilities that a standard normal variate is between 0 and Z (the shaded area).



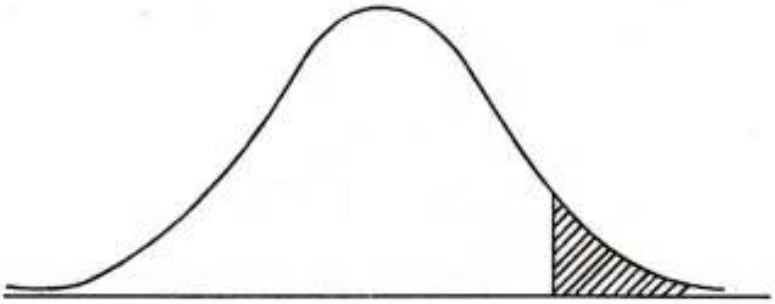
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1519	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.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	.3665	.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	.4878	.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	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.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

VIII. CRITICAL VALUES OF CHI-SQUARE (χ^2)

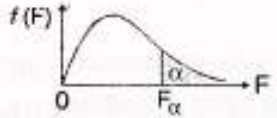


α / *d.f.*	.100	.050	.025	.010	.005	.001
1	2.71	3.84	5.02	6.63	7.88	10.8
2	4.61	5.99	7.38	9.21	10.6	13.8
3	6.25	7.81	9.35	11.3	12.8	16.3
4	7.78	9.49	11.1	13.3	14.9	18.5
5	9.24	11.1	12.8	15.1	16.7	20.5
6	10.6	12.6	14.4	16.8	18.5	22.5
7	12.0	14.1	16.0	18.5	20.3	24.3
8	13.4	15.5	17.5	20.1	22.0	26.1
9	14.7	16.9	19.0	21.7	23.6	27.9
10	16.0	18.3	20.5	23.2	25.2	29.6
11	17.3	19.7	21.9	24.7	26.8	31.3
12	18.5	21.0	23.3	26.2	28.3	32.9
13	19.8	22.4	24.7	27.7	29.8	34.5
14	21.1	23.7	26.1	29.1	31.3	36.1
15	22.3	25.0	27.5	30.6	32.8	37.7
16	23.5	26.3	28.8	32.0	34.3	39.3
17	24.8	27.6	30.2	33.4	35.7	40.8
18	26.0	28.9	31.5	34.8	37.2	42.3
19	27.2	30.1	32.9	36.2	38.6	43.8
20	28.4	31.4	34.2	37.6	40.0	45.3
21	29.6	32.7	35.5	38.9	41.4	46.8
22	30.8	33.9	36.8	40.3	42.8	48.3
23	32.0	35.2	38.1	41.6	44.2	49.7
24	33.2	36.4	39.4	43.0	45.6	51.2
25	34.4	37.7	40.6	44.3	46.9	52.6
26	35.6	38.9	41.9	45.6	48.3	54.1
27	36.7	40.1	43.2	47.0	49.6	55.5
28	37.9	41.3	44.5	48.3	51.0	56.9
29	39.1	42.6	45.7	49.6	52.3	58.3
30	40.3	43.8	47.0	50.9	53.7	59.7
35	46.1	49.8	53.2	57.3	60.3	66.6
40	51.8	55.8	59.3	63.7	66.8	73.4
45	57.5	61.7	65.4	70.0	73.2	80.1
50	63.2	67.5	71.4	76.2	79.5	86.7
55	68.8	73.3	77.4	82.3	85.7	93.2
60	74.4	79.1	83.3	88.4	92.0	99.6
65	80.0	84.8	89.2	94.4	98.1	106.0
70	85.5	90.5	95.0	100.4	104.0	112.3
75	91.1	96.2	100.8	106.4	110.3	118.6
80	96.6	101.9	106.6	112.3	116.3	124.8
85	102.1	107.5	112.4	118.2	122.3	131.0
90	107.6	113.1	118.1	124.1	128.3	137.2
95	113.0	118.8	123.9	130.0	134.2	143.3
100	118.5	124.3	129.6	135.8	140.2	149.4

IX. CRITICAL VALUES OF t



d.f.	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
inf.	1.282	1.645	1.960	2.326	2.576



X. 5% POINTS OF FISHER'S F-DISTRIBUTION

n \ m	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	α
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	250.09	252.20	254.32
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413	19.420	19.446	19.462	19.479	19.496
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8868	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6166	8.5720	8.5265
4	7.7086	6.9443	6.5914	6.3883	6.2560	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7459	5.6878	5.0281
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8753	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.4957	4.4314	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2066	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8082	3.7398	3.6688
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5108	3.4445	3.3758	3.3043	3.2298
8	5.3177	4.4590	4.0662	3.8378	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2840	3.2184	3.1503	3.0794	3.0053	2.9276
9	5.1174	4.2565	3.8626	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0001	2.9365	2.8637	2.7872	2.7007
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.6996	2.6211	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.5705	2.4901	2.4045
12	4.7272	3.8853	3.4903	3.2502	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.4663	2.3842	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.3803	2.2966	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6021	2.5342	2.4630	2.3879	2.3082	2.2230	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4035	2.3275	2.2468	2.1601	2.0658
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.1938	2.1058	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1477	2.0584	1.9604
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1071	2.0166	1.9168
19	4.3808	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.0712	1.9796	1.8780
20	4.3513	3.4928	3.0984	2.8661	2.7100	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0391	1.9464	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3661	2.3210	2.2504	2.1757	2.0960	2.0102	1.9165	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	1.9842	1.8895	1.7831
23	4.2793	3.4221	3.0280	2.7955	2.6500	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282	2.0476	1.9605	1.8649	1.7570
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9390	1.8424	1.7331
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9192	1.8217	1.7110
26	4.2252	3.3690	2.9751	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9010	1.8027	1.6906
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9736	1.8842	1.7851	1.6717
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411	1.9586	1.8687	1.7689	1.6541
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2782	2.2229	2.1768	2.1045	2.0275	1.9446	1.8543	1.7537	1.6377
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8409	1.7396	1.6223
40	4.0848	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9245	1.8389	1.7444	1.6373	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3688	2.2540	2.1665	2.0970	2.0401	1.9926	1.9174	1.8364	1.7480	1.6491	1.5343	1.3893
120	3.9201	3.0718	2.6802	2.4472	2.2900	2.1750	2.0867	2.0164	1.9588	1.9105	1.8337	1.7505	1.6587	1.5543	1.4290	1.2539
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.4591	1.3180	1.0000

XI. 1% POINTS OF FISHER'S F-DISTRIBUTION

n \ m	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	α
1	4052.2	4999.5	5403.3	5624.6	5763.7	5859.0	5928.3	5981.6	6022.5	6055.8	6106.3	6157.3	6208.7	6260.7	6313.0	6366.0
2	98.503	99.000	99.166	99.249	99.299	99.332	99.356	99.374	99.388	99.399	99.416	99.432	99.449	99.466	99.483	99.501
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.872	26.690	26.505	26.316	26.125
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.020	13.838	13.652	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.8883	9.7222	9.5527	9.3793	9.2020	9.0204
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1016	7.9761	7.8741	7.7183	7.5590	7.3958	7.2285	7.0568	6.8801
7	12.246	9.5466	8.4513	7.8467	7.4604	7.1914	6.9928	6.8401	6.7188	6.6201	6.4691	6.3143	6.1554	5.9921	5.8236	5.6495
8	11.259	8.6491	7.5910	7.0060	6.6318	6.3707	6.1776	6.0289	5.9106	5.8143	5.6668	5.5151	5.3591	5.1981	5.0316	4.8588
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511	5.2565	5.1114	4.9621	4.8080	4.6486	4.4831	4.3105
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424	4.8492	4.7059	4.5582	4.4054	4.2469	4.0819	3.9090
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315	4.5393	4.3974	4.2509	4.0990	3.9411	3.7761	3.6025
12	9.3302	6.9266	5.9526	5.4119	5.0643	4.8206	4.6395	4.4994	4.3875	4.2961	4.1553	4.0096	3.8584	3.7008	3.5355	3.3608
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911	4.1003	3.9603	3.8154	3.6646	3.5070	3.3413	3.1654
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297	3.9394	3.8001	3.6557	3.5052	3.3476	3.1813	3.0040
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948	3.8049	3.6662	3.5222	3.3719	3.2141	3.0471	2.8684
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804	3.6909	3.5527	3.4089	3.2588	3.1007	2.9330	2.7528
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822	3.5931	3.4552	3.3117	3.1615	3.0032	2.8348	2.6530
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971	3.5082	3.3706	3.2273	3.0771	2.9185	2.7493	2.5660
19	8.1850	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225	3.4338	3.2965	3.1533	3.0031	2.8442	2.6742	2.4893
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567	3.3682	3.2311	3.0880	2.9377	2.7785	2.6077	2.4212
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981	3.3098	3.1729	3.0299	2.8796	2.7200	2.5484	2.3603
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458	3.2576	3.1209	2.9709	2.8274	2.6675	2.4951	2.3055
23	7.8811	5.6637	4.7649	4.2635	3.9392	3.7102	3.5390	3.4057	3.2986	3.2106	3.0740	2.9311	2.7805	2.6202	2.4471	2.2559
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560	3.1681	3.0316	2.8887	2.7380	2.5773	2.4035	2.2107
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172	3.1294	2.9931	2.8502	2.6993	2.5383	2.3637	2.1694
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818	3.0941	2.9579	2.8150	2.6640	2.5026	2.3273	2.1315
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494	3.0618	2.9256	2.7827	2.6316	2.4699	2.2938	2.0965
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195	3.0320	2.8959	2.7530	2.6017	2.4397	2.2629	2.0642
29	7.5976	5.4205	4.5378	4.0449	3.7254	3.4995	3.3302	3.1982	3.0920	3.0045	2.8685	2.7256	2.5742	2.4118	2.2344	2.0342
30	7.5625	5.3904	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665	2.9791	2.8431	2.7002	2.5487	2.3860	2.2079	2.0062
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876	2.8005	2.6648	2.5216	2.3689	2.2034	2.0194	1.8047
60	7.0771	4.9774	4.1259	3.6491	3.3389	3.1187	2.9530	2.8233	2.7185	2.6318	2.4961	2.3523	2.1978	2.0285	1.8363	1.6006
120	6.8510	4.7865	3.9493	3.4796	3.1735	2.9559	2.7918	2.6629	2.5586	2.4721	2.3363	2.1915	2.0346	1.8600	1.6557	1.3805
∞	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073	2.3209	2.1848	2.0385	1.8783	1.6964	1.4730	1.0000

For $m > 10$ interpolate using $60/m$.

For $n > 30$ interpolate using $120/n$.

XII. FACTORS USEFUL IN THE CONSTRUCTION OF CONTROL CHARTS

	Mean-chart								*Range-chart*				
Sample size	*Factors for control limit*			*Factors for central line*	*Factors for control limit*				*Factors for central line*	*Factors for control limit*			
n	A	A_1	A_2	c_2	B_1	B_2	B_3	B_4	d_2	D_1	D_2	D_3	D_4
2	2.121	3.760	1.881	0.6642	0	1.843	0	3.267	1.128	0	3.686	0	3.267
3	1.732	3.394	1.023	0.7236	0	1.858	0	2.566	1.693	0	4.358	0	2.575
4	1.500	2.880	0.729	0.7979	0	1.808	0	2.269	2.059	0	4.698	0	2.282
5	1.342	1.596	0.577	0.8407	0	1.756	0	2.089	2.326	0	4.918	0	2.115
6	1.225	1.410	0.483	0.8686	0.026	1.711	0.030	1.970	2.534	0	5.078	0	2.004
7	1.134	1.277	0.419	0.8882	0.105	1.672	0.118	1.888	2.704	2.205	5.203	0.076	1.924
8	1.061	1.175	0.073	0.9027	0.167	1.638	0.185	1.815	2.847	0.387	5.307	0.136	1.864
9	1.000	1.094	0.037	0.9139	0.219	1.609	0.239	1.761	2.970	0.546	5.394	0.184	1.816
10	0.949	1.028	0.308	0.9227	0.262	1.584	0.284	1.716	3.078	0.687	5.469	0.223	1.777
11	0.905	0.973	0.285	0.9300	0.299	1.561	0.321	1.679	3.173	0.812	5.534	0.256	1.744
12	0.866	0.925	0.256	0.9359	0.331	1.541	0.354	1.646	3.258	0.924	5.592	0.284	1.716
13	0.832	0.883	0.249	0.9410	0.359	1.523	0.382	1.618	3.336	1.026	5.646	0.308	1.692
14	0.802	0.848	0.235	0.9453	0.384	1.507	0.406	1.594	3.407	1.121	5.693	0.329	1.671
15	0.775	0.816	0.223	0.9490	0.406	1.492	0.428	1.572	3.472	1.207	5.737	0.348	1.652
16	0.750	0.788	0.212	0.9523	0.427	1.478	0.448	1.542	3.352	1.285	5.279	0.365	1.636
17	0.728	0.762	0.203	0.9551	0.445	1.465	0.466	1.534	3.588	1.359	5.817	0.379	1.621
18	0.707	0.738	0.816	0.9576	0.461	1.454	0.482	1.518	3.640	1.426	5.854	0.404	1.608
19	0.688	0.717	0.187	0.9599	0.477	1.443	0.497	1.503	3.689	1.490	5.888	0.404	1.596
20	0.671	0.697	0.180	0.9619	0.491	1.433	0.510	1.490	3.735	1.548	5.922	0.414	1.585
21	0.655	0.670	0.173	0.9638	0.504	1.424	0.523	1.447	3.778	1.606	5.950	0.425	1.575
22	0.640	0.662	0.167	0.9655	0.516	1.415	0.534	1.466	3.819	1.659	5.979	0.434	1.566
23	0.626	0.647	0.162	0.9670	0.527	1.407	0.545	1.455	3.858	1.710	6.006	0.443	1.557
24	0.612	0.632	0.157	0.9684	0.538	1.399	0.555	1.445	3.895	1.759	6.031	0.452	1.548
25	0.600	0.319	0.153	0.9696	0.548	1.392	0.565	1.435	3.931	1.804	6.058	0.459	1.541

XIV. RANDOM NUMBERS

58941	72711	39408	91620	27963	96478	21559	19246	88097	44026
02349	71389	45608	60947	60775	73181	43264	56895	04232	59604
89210	44546	27174	27499	53523	63110	57106	20865	91683	80688
11826	91326	29664	01603	23156	89223	43429	95353	44662	59433
69810	17100	35066	00815	01552	06392	31437	70385	45863	75971
81060	33449	68055	83844	90942	74857	52418	68723	47830	63010
56135	80647	51404	06626	10042	93629	37609	57215	08409	81906
57361	65304	93258	56760	63348	24949	11839	29793	37457	59377
24548	56415	61927	64416	29934	00755	09418	14230	62887	92683
66504	02036	02922	63569	17906	38076	32135	19096	96970	75917
45068	05520	56321	22693	35089	07694	04252	23791	60249	83010
99717	01542	72990	43413	59744	44595	71326	91382	45114	20245
05394	61840	83089	09224	78530	33996	49965	04851	18280	14039
38155	42661	02363	67625	34683	95372	74733	63558	09665	22610
04319	04318	99387	86874	12549	38369	54952	91579	26023	81076
18134	90062	10761	54548	49505	52685	63903	13193	33905	66936
32012	42710	34650	73236	66167	21788	03581	40699	10396	81827
78101	44392	53767	15220	66319	72953	14071	59148	95154	72852
23469	42846	94810	16151	08029	50554	03891	38313	34016	18671
35342	56119	97190	43635	84249	61254	80993	55431	90793	62603
65846	18076	12415	30193	42777	85611	57635	51362	79907	77364
22184	33998	87436	37430	45246	11400	20986	43996	73112	88474
83668	66236	79665	88312	93047	12088	86937	70794	01041	74867
90083	70696	13558	98995	58159	04700	90443	13168	31553	67891
97765	27552	49617	51734	20819	70198	67906	00880	82899	66065
49988	13176	94219	88698	41755	56216	66832	17748	04963	54859
78257	86249	46134	51865	09836	73966	65711	41699	11732	17173
30946	22210	79302	40300	08852	27528	84648	79589	95295	72895
19468	76358	69203	02760	28625	70476	76410	32988	10194	94917
30806	80857	84383	78450	26245	91763	73117	33047	03577	62599
42163	68332	98851	50252	56911	62693	73817	98693	18728	94741
39249	51463	95963	07929	66728	47761	81472	44806	15592	71357
88717	29289	77360	09030	39605	87507	85446	51257	89555	75520
16767	57345	42285	56670	88445	85799	76200	21795	38894	58070
77516	96648	51868	48140	13583	94911	13318	64741	64336	95103
87192	66483	55649	36764	86132	12463	28385	94242	32065	45233
74078	64120	04643	14351	71381	26133	68269	65145	28152	39087
94119	20108	78101	81276	00835	63835	87174	42446	08882	27067
62180	27453	18567	55524	86088	00069	59254	24654	77371	26409
56199	05993	71201	78852	65889	32719	13758	23937	90740	16866
04994	09879	70337	11861	69032	51915	23510	32050	52052	24004
21725	43827	78862	67699	01009	07050	73324	06732	27510	33761
24365	37661	18956	50064	39500	17450	18030	63124	48061	59412
14762	69734	89150	93126	17700	94400	76075	08317	27324	72723
28387	99781	52977	01657	92602	41043	05686	15650	29970	95877

Source: Extracted from Table of 105,00 Random Decimal Digits Statement No. 4914, File No. 261-A (Washington D.C. Interstate Commerce Commission, 1949)

Financial Management

Principles and Practice

Dr. S.N. MAHESHWARI, Ph.D.

9th Revised Edn. Knowledge-packed pages *xx* + 1510 ISBN 81-8054-158-4

Contents

International Financial Management

An Indian Perspective

Dr. R.L. VARSHNEY* and Dr. S. BHASHYAM**

Third Edn. Knowledge-packed pages *xvi* + 546 ISBN 81-8054-305-6

Though there are a number of books on the subject, both by foreign and Indian authors, no book is comprehensive enough to meet the requirements of Indian students. An attempt has now been made to fill the gaps.

Contents

* Dr. R.L. Varshney, former Director, Indian Institute of Foreign Trade, has taught International Finance in the Indian Institute of Management, Lucknow, and a number of other management institutes.

** Dr. S. Bhashyam, Professor, International Finance & Liquidity, at the University of Delhi, has been the Coordinator of the Master's Programme in International Business and has also been teaching International Finance in a number of management institutes.

Foreign Exchange and Risk Management

C. JEEVANANDAM

PSG Institute of Management, Coimbatore

Ninth Edn. Reprint Pages 450 ISBN 81-8054-155-x

The book blends sound theoretical knowledge of foreign exchange economics with practical and procedural aspects of banks and other institutions connected with foreign exchange. Adequate coverage has also been provided for risk management by banks as well as exporters and importers. Each topic is comprehensively dealt with presents in a cogent and understandable manner materials that lie scattered and sometimes not easily accessible.

The book meets the requirements of post-graduate courses in commerce and economics and MBA for papers such as 'Foreign Exchange", 'Foreign Exchange Management', 'Exchange Risk Management', and 'International Finance'.

Human Resource Development

Dr. P.C. TRIPATHI, Ph.D.
Formerly Prof. and Head of the Deptt. of Business Administration
Sukhadia University, Udaipur

Fourth Revised Edn. Reprint Pages *xvi* + 396 ISBN 81-7014-992-4

This book describes in detail the theory and practice of HRD/HRM. The book is specially designed to serve as a textbook for this paper at the post-graduate level in various Indian Universities.

Contents

Personnel Management & Industrial Relations

Dr. P.C. TRIPATHI, Ph.D.

17th Revised Edn. Reprint Pp. *xvi* + 552 ISBN 81-8054-135-5

A comprehensive and cogent presentation of the subject in the light of the practices prevailing in our country.

Contents

Organisational Development and Human Resource Development

Dr. P.C. TRIPATHI

First Edition Pages *xvi* + 356 ISBN 81-8054-074-X

Contents

Financial Analysis and Financial Management

A Contemporary Approach

R.P. RUSTAGI, M. Com., M.Phil., F.C.S.
Shri Ram College of Commerce
University of Delhi

First Edition | Pages : *xx* + 948 | ISBN 81 – 8054–281–5

'Financial Analysis and Financial Management' has been prepared to meet the requirements of students taking CA (Final) and other higher level courses in Finance. The book presents an analytical framework of the related subject-matter.

Special Features

- Concepts and procedures have been explained in a well-knit manner.
- Sufficient **examples** have been provided for a better grasp.
- 398 practical problems have been given with solutions in the form of ***Graded Illustrations.***
- ***Solutions to latest question papers*** set at the CA (Final) examinations are provided.
- ***Points To Remember*** at the end of the Text present each chapter in a capsule form.
- ***Self Review Assignments*** contain short concept questions and essay type questions.
- ***Eight Model Test Papers with answers*** have been given in the book.

Contents

SULTAN CHAND & SONS, *Educational Publishers*

Marketing Management

An Indian Perspective

Dr. R.L. VARSHNEY* and Dr. S.L. GUPTA**

Third Revised Edn. | Pp. *xxiv* + 1126 | ISBN 81-7014-318-8

Written in a lucid style, the book has more than a thousand Indian examples at appropriate places. It also contains a number of Indian cases and live situations.

Contents

* Dr. R.L. Varshney, Former Director, Indian Institute of Foreign Trade, has been teaching in various Management Institutes for over 40 years.

** Dr. S.L. Gupta, after working for 5 years with the Industry, has been teaching in the Apeejay School of Marketing, New Delhi since 1992.

Marketing Management

Dr. C.B. GUPTA
Reader in Commerce
Shri Ram College of Commerce
University of Delhi

Dr. N. RAJAN NAIR
Professor & Head
Deptt. of Rural Marketing Management
Kerala Agricultural University

Seventh Edition | Pages *viii* + 608 | ISBN 81-7104-961-4

Contents

Advertising and Sales Promotion

Dr. S.L. GUPTA • Dr. V.V. RATNA

First Edn. | Pp. *xvi* + 648 | ISBN 81-8054-111-8

Contents

Consumer Behaviour

An Indian Perspective

Dr. S.L. GUPTA* • SUMITRA PAL**

First Edn. Reprint | Pages *xvi* + 556 | ISBN 81-7014-795-6

- This book provides insight into the consumer behaviour with focus on Indian environment.
- The contents will enable the marketers to apply the concepts to real time marketing.

Contents

Understanding Consumer Behaviour • Consumer Research • Market Segmentation • Consumer Needs and Motivation • Consumer Personality • Consumer Perception • The Process of Learning an Consumer Behaviour • The Nature of Consumer Attitudes • Models of Consumer Behaviour • Group Dynamics and Consumer Reference Groups • Communication, Advertising and Consumer Buying Behaviour • The Family and Life Style Marketing • Social Class and Consumer Behaviour • Culture, Sub-Culture and Cross Culture • The Process of Innovations and Diffusion of Innovation • Consumer Behaviour as a Decision Process Maintaining Consumer Satisfaction • Consumerism and Public Policy Issues, Organisational Buyer Behaviour • Case Study--Appendices • Glossary.

* Dr. S.L. Gupta is an Associate Professor of Marketing at Appeejay School of Marketing. Dr. Gupta possesses 4 years' corporate experience and 7 years' academic experience and specialises in marketing stream. He has been teaching Marketing Management, Marketing Research, Sales and Distribution Management, Consumer Behaviour at various Institutes since 1992. He is accredited management teacher from All India Management Association.

** Ms. Sumitra Pal has extensive industrial and teaching experience in various management schools.

International Marketing Management

An Indian Perspective

Dr. R.L. VARSHNEY, M.Com., Ph.D.
Formerly Director, Indian Institute of Foreign Trade, New Delhi

B. BHATTACHARYA, M.A.
Dean
Indian Institute of Foreign Trade, New Delhi

XVIII Revised & Enlarged Edn. | Pp. *xvi* + 597 | 22 × 14 cm | ISBN 81-8054-223-8

It deals with Why, When, What, Where and How of export marketing.

Special Features

- All data, developments and policies, both national and international, have been brought up-to-date.
- This book introduces a number of case histories and cases.

Contents

PART I— International Trading Environment (Pp. 131)

Framework of International Marketing • Basis of International Trade • Recent Trends in World Trade • Foreign Trade and Economic Growth • Balance of Payments and Instruments of Trade Policy • International Economic Institutions • Regional Economic Groupings.

PART II—India's Foreign Trade (Pp. 90)

Recent Trends in India's Foreign Trade • Institutional Infrastructure for Export Promotion in India • India's Trade Policy • Export Assistance.

PART III—International Marketing (Pp. 296)

Identifying Foreign Markets • Product Planning for Export • Pricing for Exports • Market Entry and Overseas Distribution System • Distribution Logistics for Exports • Promoting Products Internationally • Overseas Market Research • Marketing Plan for Exports • Decision-making Framework for Export Operation • New Techniques in International Marketing • Terms of Payments and Export Finance • Management of Risks in International Marketing • Project and Consultancy Exports • Global Marketing of Services • Multinationals : Their Role in International Marketing • State Trading in India • Legal Dimensions of International Marketing • Export Documents and Procedure.

APPENDICES (Pp. 62)

Cases • Selected Sources of Information • Suggested Readings • Review Questions.

Marketing Research

Principles, Applications and Cases

Dr. D.D. SHARMA

Technical Teachers' Training Institute, Chandigarh

Formerly Associate Professor, Deptt. of Business Management

Punjab Agricultural University, Ludhiana

2nd Edn. Reprint | Knowledge-packed pages *xvi* + 552 | ISBN 81-7014-658-5

This book provides a down-to-earth description of techniques involved in designing, conducting and applying marketing research to the problems in business organisations.

- The emphasis is on developing an understanding of the principles and their applications.
- Case studies on actual Indian market situations have been included.

This book can be used as a textbook by the management students specialising in the area of marketing.

Contents

A Textbook of

Research Methodology in Social Sciences

Dr. P.C. TRIPATHI, Ph.D.

Formerly Head, Deptt. of Business Administration

Sukhadia University, Udaipur

5th Revised Edn. | Knowledge packed pages *xvi* + 368 | ISBN 81-8054-296-3

The book has been written specifically to meet the needs of students and researchers in social sciences. It covers the syllabi of theory of research methodology paper of various universities. This book is written strictly as a textbook.

Contents

Managerial Economics

Dr. R.L. VARSHNEY, Ph.D.
Former Director, Indian Institute of Foreign Trade, New Delhi

Dr. K.L. MAHESHWARI, Ph.D.
Professor of Applied Economics, Lucknow University, Lucknow

18th Revised Edn. Reprint Pp. *xx* + 836 ISBN 81-8054-148-7 23 × 14 cm

It is meant for students of M.Com. and Business Management Courses and Business Managers.

A concerted effort has been made to impart empirical content or practice-orientation to the various concepts of Pure Economics. The book contains many decision-making situations in the form of Illustrations.

Contents

Economic Environment of Business

Theory and the Indian Case

Dr. M. ADHIKARY

Director Emeritus, New Delhi Institute of Management
Ex-Dean, FMS, DU, Professor and Management Consultant
Ex-Director, Shriram Research Centre

10th Rev. Edn. Reprint 22 × 14 cm. Pp. *xvi* + 768 ISBN 81-8054-239-4

The book is primarily addressed to the students of M.B.A., M.A. (Economics, Business Economics), I.E.S., I.I.M.A., M.Com., and M.Phil. It is also expected to be of immense help to teachers, business executives, professional managers, corporate planners and government policy-makers.

About the Book

Business is an economic activity; business decision-making is an economic process. It is, therefore, important to identify and understand the critical elements of the economic environment of business.

The main purpose of this book is to build up a few macro-economic concepts and theories into an *analytical framework* with reference to which one can attempt a meaningful evaluation of the economic environment of business in India.

The challenges before Indian management are thus objectively reworked in the context of an analysis of business problems and prospects of the Indian economy of today.

For ready reference, the relevant up-to-date statistical information about the Indian economy has been put together in the *Appendix*.

The present edition is thoroughly revised in terms of the empirical contents, covering latest developments. In particular, additions/alterations/revisions undertaken in the former edition are :

- Focus on privatisation and emerging market-friendly approach to competitive environment.
- Up-to-date examples and footnotes in view of recent developments in Indian corporate sector.
- Supply side economic principles and policy implications.
- Additional diagrams and models to explain inflation, stagflation, etc.
- New sections on Infrastructure Sector, Social Sector, Bureaucracy and Business, Economic Offences in India, etc.
- Latest economic policy statements on monetary, fiscal and physical fronts; Minimum Economic Programme of the new Government, Latest Pay Commission, Updated Economic Survey, Future of Economic Reform in India, etc.
- Absolutely new chapter on Current Trends and Tendencies.
- Completely revised Data-Environment, incorporating new data, charts and statistical analysis.
- The book projects a viewpoint which is refreshing as well as thought-provoking.

Contents

About the Author

Manabendra Adhikary (born 1941), Ph.D. (Bloomington, Indiana). A.E.A. Diploma (Boulder, Colorado), M.A. (Delhi School of Economics) is Ex-Dean, Faculty of Management Studies, University of Delhi. Dr. Adhikary is having more than 35 years of teaching, research and consultancy experience at home and abroad.

Managerial Economics

Dr. P.L. MEHTA
Head, Department of Economics,
Shri Ram College of Commerce, University of Delhi,
Formerly, Asstt. Professor, Indian Institute of Technology, Delhi

10th Revised & Enlarged Edn. Reprint Pp. *xx* + 838 ISBN 81-8054-262-9

This very popular textbook is designed for MBA, M.A. (Business Economics), M.Com., M.A. (International Business), AIMA, M.F.C., PGDBA, B.B.A and B.B.M. courses of Indian Universities and Institutes.

Distinguishing Features

- Very comprehensive text with lucid and easy language.
- Emphasis given to the application of the analytical and empirical tools.
- Greater emphasis on explanation of the more difficult concepts and methods.
- Large number of Case Studies alongwith the solved examples in the text.
- Summary, meaning of Important concepts given at the end of each chapter as well as more than 800 questions, problems and review questions.

Contents

PART I : Meaning and Scope • Fundamental Concepts, Models and Methods • Alternative Objectives of the Firm.

PART II : Theory of Consumer Behaviour • Demand Analysis • Elasticity of Demand and Demand Estimation • Demand Forecasting—An Introduction • Methods of Demand Forecasting • Advertising and Sales Promotion.

PART III : Supply and Production Decisions • Cost of Production • Inventory Cost Management.

PART IV : Theory of Pricing—Perfect Competition and Monopoly • Theory of Pricing—Monopolistic Competition, Duopoly and Oligopoly • Pricing Practices and Strategies • Advanced Topics in Pricing Theory • General Conditions in Pricing and Pricing Forecasting • Factor Markets and Factor Prices.

PART V : Profit—Theory and Measurement • Profit : Policy, Planning, Control and Forecasting.

PART VI : Capital Budgeting : Evaluating Capital Projects • Capital Budgeting : Cost of Capital • Decision Analysis • Risk in Project Analysis • The Location Decisions.

PART VII : National Income and Business Cycles • Role of Government in Market Economy • Government and Business (An Analysis of Economic Policy) • Public Sector Decisions.

PART VIII : International Trade and International Finance • Economic Decisions in Multinational Setting.

PART IX : Linear Programming • Input-Output Analysis • Game Theory.

Appendices : Integrating Case Study • Review Questions • Answers to Questions • Present Value Tables • Mathematics for Managerial Economics.

Comprehensive Managerial Economics

(As per the UGC proposed syllabus for MBA)

Dr. P.L. MEHTA

Pages *xvi* + 660 ISBN 81-8054-140-1

It is an exhaustive book written in the light of the latest syllabus proposed by the University Grants Commission for MBA coursework in Indian universities.
The book covers all the three aspects of the subject—the theory, problems and cases. It provides a unified and comprehensive view of the subject matter of Managerial Economics.

Contents

Meaning and Scope • Fundamental Concepts, Models and Methods • Alternative Objectives of the Firm • Theory of Consumer Behaviour • Demand Analysis • Elasticity of Demand and Demand Estimation • Demand Forecasting—An Introduction • Methods of Demand Forecasting • Advertising & Sales Promotion • Supply and Production Decision • Cost of Production • Inventory Cost Management • Theory of Pricing—Perfect Competition and Monopoly of the Limiting Cases. Theory of Pricing—Monopolistic Competition, Duopoly and Oligopoly • Pricing practices & Strategies • Advanced Topics in Pricing Theory • General Considerations in Price Forecasting • Profit—Theory and Measurement • Profit : Policy, Planning, Control and Forecasting (Break-even Analysis) • Business and Macro Economic Environment • National Income—Concepts and Measurement • Consumption Function and Multiplier • Fiscal Policy and Income Determination • The Level of Investment • Demand and Supply of Money • Monetary and Real Sectors Taken Together • Central Banking and Monetary Controls—Business Cycles • Inflation & Deflation.
Appendices : Interrating case studies • Review Questions • Answers to Questions • Mathematical Treatment of Managerial Economics.

Business Environment

Dr. P.K. GHOSH

Formerly Professor of Commerce, University of Delhi, Delhi

Contents

Changing Perspective of Business in India : An Overview • Dimensions of Business Environment : Dynamics and Specificity. • Need for and Importance of Environmental Analysis • Macro-economic Environment : Structural Adjustment Programme • Fiscal and Monetary Policies • Industrial Policy Changes • Public Sector : Performance and Disinvestment • Privatisation—Why and How ? • Small-Scale Industries : Policy Issues • Economic Concentration : Relevance of Legal Regulation • Money Market and Banking Sector Development • Service Sector Reforms and Regulation : Power Supply, Telecom and Insurance • Foreign Investments and Collaboration • Securities Markets : SCRA and SEBI Act • Globalisation : Implications and Impact • Trade Policy Reforms : India and WTO • Consumerism and Consumer Protection • Policy Measures on Environmental Protection • The Problem of Industrial Sickness : Policy Frame • Economic Showdown and Corporate Response • APPENDIX : Case Studies.

The Indian Economy : Environment and Policy

I.C. DHINGRA

Reader in Economics, Bhagat Singh College,
University of Delhi, Delhi

19th Thoroughly Revised Edn. 20 × 30 cm. Pp. *xx* + 756 Chapters 30 ISBN 81-8054-280-7

Distinguishing Features

- The emphasis is on analytical treatment.
- All the relevant up-to-date facts and figures culled from authentic sources have been used to make analytical study of different aspects of the economy.
- Various problems of the Indian economy have been analysed with proper theoretical backdrop in their relevant context.
- The economic scene has been presented in a very simple language and easy-to-understand style.
- The useful data have been presented in the form of small concise tables.
- Graphs are included.
- A virtual goldmine of informed analysis of problems of Indian economy.

Indian Financial System

P.N. VARSHNEY
Former Professor and Head,
Deptt. of Business Economics,
University of Delhi

D.K. MITTAL
Reader, Deptt. of Commerce,
Shri Ram College of Commerce,
University of Delhi

Sixth Revised Edition ISBN 81-8054-274-2 Pp. *xxxix* + 724

Contents

Business Statistics

Dr. S.P. GUPTA
Dean, Faculty of Management Studies, University of Delhi, Delhi
Dr. M.P. GUPTA
Formerly Dean, Faculty of Management Studies, University of Delhi, Delhi

14th Revised Edn. | Pp. *xii*+688 | Soft cover | 479 Solved Illustrations
ISBN 81-8054-327-7 | 891 Exercises with answers

Contents

Business Statistics—What and Why • Collection of Data • Presentation of Data • Measures of Central Tendency • Measures of Variation • Skewness, Moments and Kurtosis • Correlation Analysis • Regression Analysis • Index Numbers : Concepts and Applications • Business Forecasting and Time Series Analysis • Probability • Probability Distributions • Sampling and Sampling Distributions • Estimation of Parameters • Tests of Hypothesis • Small Sampling Theory • Chi-Square Test • Analysis of Variance • Statistical Quality Control • Partial and Multiple Correlation and Regression • Statistical Decision Theory • Appendix.

Business Mathematics

Dr. D.C. SANCHETI, Ph.D.
Formerly, Joint Director of Studies, Institute of Chartered Accountants, New Delhi.
V.K. KAPOOR, M.A.
Sri Ram College of Commerce, University of Delhi, Delhi

11th Edn. Reprint | Pp. *xviii* + 1179 | 22 × 14 cm. Soft Cover | ISBN 81-7014-121-4

The chief merit of the book is its simplicity. The book has 989 illustrations, 112 diagrams and 1149 exercises with answers for practice.

Contents

Logical Statements and Truth Tables • Theory of Sets • Boolean Algebra • Real Number Systems • Groups, Ring and Field • Indices and Surds • Logarithms • Equations : Linear Quadratic, Cubic and Higher Order • Permutations and Combinations • Binomial Theorem • Mathematical Induction, Sequence and Series • Arithmetic and Geometric Progressions • Convergence and Divergence of Series • Circular Functions and Trigonometry • Coordinate Geometry • Functions, Limits and Continuity • Differential Calculus • Integral Calculus • Vector Algebra • Matrix Algebra. Applications to Business and Economics • Linear Programming • Probability • Some Additional Topics • Tables • Index.

Quantitative Methods

Dr. D.C. SANCHETI, Ph. D.
V.K. KAPOOR, M.A. Dr. P.L. MEHTA, Ph.D.
Shri Ram College of Commerce, University of Delhi, Delhi

IV Thoroughly Revised Edn. Pp. *xviii* + 1052 634 Solved Illustrations 22 × 14 cm

Contents

Section—A : MATHEMATICS (Pp. 284, Solved Examples 175, Unsolved Probs. 253) : Matrix Algebra. Functions, Limits and Continuity • Differential Calculus • Integral Calculus • Linear Programming.

Section—B : STATISTICS (Pp. 444, Solved Examples 163, Unsolved Probs. 557) : Correlation Analysis • Regression Analysis • Probability and Expected Value • Statistical Decision Theory • Sampling and Designing of a Sample Survey • Test of Hypothesis • Chi-Square Test

Section—C : ECONOMIC TECHNIQUES (Pp. 374, Solved Probs. 150, Unsolved Probs. 193) : The Demand • Consumer Behaviour • Elasticity of Demand • Demand Forecasting Introduction • Input-Output Analysis • Time Series • Index Numbers • Multiple and Partial Correlation and Regression • Empirical Production Function Cost Analysis • Single and Multi-Product Firms Factor Demand and Decisions Under Uncertainty • Scanner.

Operations Research

Dr. KANTI SWARUP, Ph.D.
Formerly Professor, Indian Institute of Public Administration, New Delhi

Dr. P.K. GUPTA, Ph.D. *J.V. Jain College, Saharanpur*
Dr. MAN MOHAN, M.Sc., Ph.D. *Ramjas College, University of Delhi, Delhi*

12th Revised Edn. Pp. *xviii* + 828 2 × 14 cm. ISBN 81-8054-226-2
300 Solved Examples 220 Diagrams Over 900 Unsolved Problems with Answers

Special Features

- It is designed to satisfy the long-felt need of students of O.R., Business Systems Analysis, Management, Engineering, Mathematics, Statistics, Commerce and executives.
- It is rigorous in its treatment of the theory, comprehensive and lucid in its explanation of techniques.
- All important Algorithms have been summarised in a step-wise manner, followed by their Flow Charts.
- Great emphasis on mathematical formulation of O.R. problems through sample problems arising in the fields of Management, Economics, Defence, Manpower Planning, Agriculture, etc.
- Addition of two new chapters on Decision Analysis and Resource Analysis in Network Scheduling.
- A chapter on Case Studies for the students of Management.
- An Appendix presents Answers to Problems contained in the book.

Problems in Operations Research

Dr. P.K. GUPTA, Ph.D. *J.V. Jain College, Saharanpur*
Dr. MAN MOHAN, M.Sc., Ph.D. *Ramjas College, University of Delhi, Delhi*

10th Edn. Pages *xvi* + 772 Chapters 30 ISBN 81-8054-030-8
760 Typical Problems Fully Solved 350 Unsolved Problems with Answers

Special Features

- Contains sufficiently large number of solved problems on each topic.
- Problems have been framed so as to include ticklish points.

Contents

Sultan Chand & Sons
A business with a cause
23, Daryaganj, New Delhi-110 002
Phones : 23266105, 23277843, 23281876, 23286788; Fax : 011-2326-6357